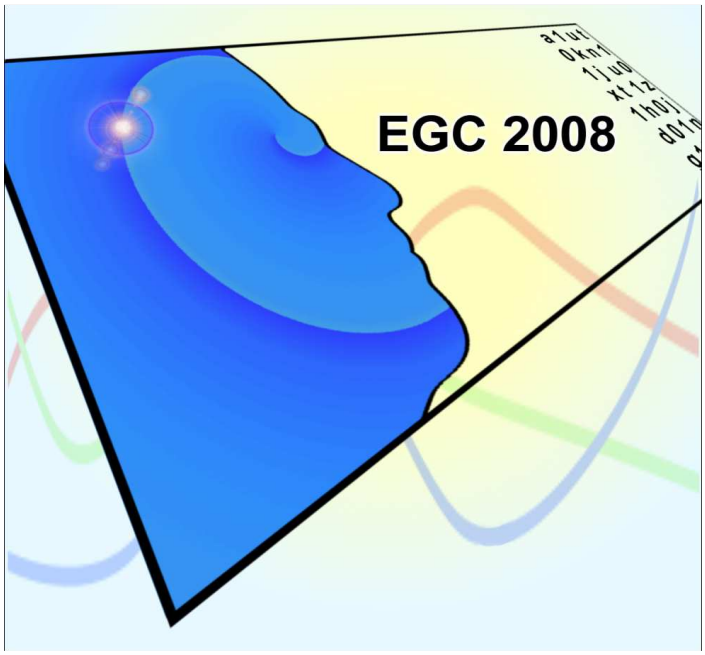
8èmes Journées Francophones
Extraction et Gestion des Connaissances
Sophia Antipolis 29 janvier 2008

Atelier



Modélisation utilisateur et personnalisation d'interfaces Web

Organisateurs :

- Zeina Jrad (INRIA)
- Abdouroihamane Anli (INRIA)
- Marie-Aude Aufaure (SUPELEC & INRIA)

Responsables des Ateliers EGC :

Alzennyr Da Silva (INRIA, Rocquencourt)
Alice Marascu (INRIA, Sophia Antipolis)
Florent Masseglia (INRIA, Sophia Antipolis)

http://www-sop.inria.fr/axis/egc08

EGC

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE
INRIA
centre de recherche SOPHIA ANTIPOLIS - MÉDITERRANÉE

Objectifs du Workshop

Au cours de ces dernières années, de nombreux travaux se sont intéressés à la conception d'interfaces Web sensibles au contexte d'interaction, afin de les rendre plus adaptatives, plus personnalisées. Les éléments de contexte peuvent inclure, par exemple, des informations relatives à l'utilisateur (buts, préférences, historique des navigations, etc.), l'environnement (lieu, luminosité, bruit, temps, etc.), la plate-forme d'interaction (téléphone portable, PDA, PC, etc.) ou encore toute information pertinente pouvant être utilisée pour caractériser les conditions d'interaction. L'acquisition, la modélisation et le traitement de ces contextes d'interaction jouent un rôle fondamental dans le développement des systèmes adaptatifs en général et des applications web personnalisées, en particulier.

L'atelier Modélisation Utilisateur et Personnalisation d'interfaces Web a pour objectif de rassembler les chercheurs intéressés par la problématique de l'extraction, la modélisation et la gestion des connaissances utilisateur pour l'adaptation et la personnalisation de l'interaction dans les applications web. Le but est de faire le point sur la prise en compte des connaissances d'interaction et de discuter des problèmes génériques, des outils et des méthodes employées.

Les articles du workshop présentent des travaux menés dans différents domaines et reliés à des thèmes variés :

- Le premier thème concerne la modélisation contextuelle. Gensel et al. présentent les travaux menés sur la définition et l'exploitation de modèles de contexte, dans le cadre des Systèmes d'Information sur le Web (SIW) à usage collaboratif à des fins d'adaptation. Jrad et al. présentent un modèle qui représente l'utilisateur et son contexte de navigation dans un portail dédié au tourisme électronique.
- Le second thème concerne le Web Usage Mining. Mroué et al. décrivent une approche pour rechercher automatiquement les comportements prototypiques dans un ensemble de parcours recueillis pour un site web. Vuillemot propose une interface de visualisation personnalisée d'un site Web regroupant des conférences scientifiques de manière structurée sous forme d'un corpus. Cette interface est multi points de vue et combine entre autres lieux, dates et mots clé de manière synchronisée. Ouamani et al. décrivent la conception d'un système multi-agent, PWUM, dédié au web usage mining pour la personnalisation du web.
- Enfin, les travaux de Dominguès sont reliés à l'exploitation de commentaires linguistiques dans un contexte de conception de carte sur mesure. L'article présente une expérimentation mise en place auprès d'utilisateurs afin de recueillir des commentaires sur des cartes géographiques et de les exploiter pour aider l'utilisateur à concevoir une carte adaptée à son besoin.

Comités

Comité d'organisation

Zeina Jrad (INRIA Paris-Rocquencourt, projet Axis)
Abdouroihamane Anli (INRIA Paris-Rocquencourt, projet Axis)
Marie-Aude Aufaure (SUPELEC et INRIA Paris-Rocquencourt, projet Axis)

Comité de programme

Yves Lechevallier (INRIA Paris-Rocquencourt)
Brigitte Trousse (INRIA Sophia-Antipolis)
Bernard Senach (INRIA Sophia-Antipolis)
Emmanuelle Grislin-Le Strugeon (LAMIH – Université de Valenciennes)
Sophie Lepreux (LAMIH – Université de Valenciennes)
Bénédicte Le Grand (Lip6 – Université Paris6)
Thierry Artières (Lip6 – Université Paris6)
Patrick Gallinari (Lip6 – Université Paris6)
Charles Tijus (Université de Paris8)
Nabil Layaida (INRIA Grenoble - Rhône-Alpes)

Contact : {Zeina.Jrad, Abdouroihamane.Anli, marie-aude.aufaure}@inria.fr

Table des matières

Modèles de contexte pour l'adaptation à l'utilisateur dans des Systèmes d'Information Web collaboratifs

Jérôme Gensel*, Marlène Villanova-Oliver*,
Manuele Kirsch-Pinheiro**

*Laboratoire d'Informatique de Grenoble, BP72, 38402 Saint Martin d'Hères cedex
Prénom.Nom@imag.fr
http://lig.imag.fr

** Department of Computer Science - Katholieke Universiteit Leuven, Leuven, Belgium
Manuele.KirschPinheiro@cs.kuleuven.be

Résumé. Les Systèmes d'Information sur le Web (SIW) permettent d'acquérir, de structurer, de stocker, de gérer, et de diffuser de l'information en s'appuyant sur une infrastructure Web. Pour ces systèmes, s'*adapter* à l'utilisateur est un défi essentiel, gage du confort de l'utilisateur, mais aussi et surtout de leur attractivité et donc de leur pérennité. L'adaptation nécessite la gestion d'un ensemble d'informations à propos de l'utilisateur (besoins, préférences, etc.) appelé *profil*. Plus générale est la notion de *contexte* qui regroupe également des informations aussi diverses que les caractéristiques matérielles et logicielles du dispositif d'accès, la qualité de service du réseau, les paramètres physiques de l'environnement, etc. Nous présentons ici les travaux que nous avons menés sur la définition et l'exploitation de modèles de contexte, dans le cadre de SIW à usage collaboratif à des fins d'adaptation.

1 Introduction

Les Systèmes d'Information sur le Web (SIW) permettent d'acquérir, de structurer, de stocker, de gérer, et de diffuser de l'information en s'appuyant sur une infrastructure Web. Un SIW offre un accès universel ou contrôlé à un espace d'informations sur lequel divers traitements sont activables, le plus souvent par le biais de requêtes. Les SIW sont de fait exploités par différents types d'application qui ont donné naissance à de nombreux domaines dont le nom est préfixé par *e-* , comme le commerce et la vente (*e-business*), l'enseignement à distance (*e-learning*), etc. Ces domaines présentent chacun des besoins spécifiques, la modélisation et les services à mettre en œuvre dans le cas d'un SIW pour le *e-commerce* sont bien différents que ceux dédiés à un SIW pour le *e-learning*, par exemple. Pour ces systèmes, s'*adapter* à l'utilisateur est un défi essentiel, gage du confort de l'utilisateur, mais aussi et surtout de leur attractivité et donc de leur pérennité. L'adaptation est le processus qui amène l'utilisateur à éprouver le sentiment que le système a été conçu spécialement pour lui. Le système doit donc gérer un ensemble d'informations à propos de l'utilisateur (besoins, préférences, etc.). On donne généralement à cet ensemble le nom de *profil*. Le profil est exploité par le système pour décider ce qui doit être présenté à l'utilisateur.

Dans les SIW, le contenu, la navigation, la présentation et les fonctionnalités sont des cibles potentielles pour l'adaptation comme l'ont montré Brusilovski (1998), Raad et Causse

(2002), Paterno et Mancini (1999), Koch (2000), et Frasincar et Houben (2002). En amont du processus (et du résultat) que représente l'adaptation, on trouve souvent la notion de *contexte* qui regroupe des informations aussi diverses que les caractéristiques matérielles et logicielles du dispositif d'accès, la qualité de service du réseau, les paramètres physiques de l'environnement, etc. Ces informations sont essentielles pour l'adaptation, notamment dans le cas de SIW ubiquitaires accédés depuis des dispositifs mobiles. En général, les contours du contexte sont flous et varient avec les besoins de l'application. Si, du point de vue de l'utilisateur, le contexte est ce qui l'entoure, du point de vue du système, une description idoine de l'utilisateur peut être considérée comme faisant partie du contexte.

Cet article traite de la représentation et de l'exploitation de modèles de contexte, à des fins d'adaptation à l'utilisateur. Il est organisé comme suit : la section 2 introduit la notion de *contexte* en rappelant les définitions, représentations, et modes d'acquisition les plus courants. La section 3 présente un modèle de contexte dédié aux Systèmes d'Information collaboratifs, exploité pour adapter une information partagée par les membres d'un groupe, dite *conscience de groupe*. Enfin, la section 4 conclut cet exposé.

2 Notion de contexte

2.1 Définitions

L'informatique sensible au contexte est apparue dans le milieu des années quatre-vingt dix impulsée par les travaux de Schilit et Theimer (1994). Ce terme fait référence à des systèmes capables de percevoir un ensemble de conditions d'utilisation – le *contexte* – afin d'adapter en conséquence leur comportement en termes de délivrance d'informations et de services (Cheverest et al. (2002), Chaari et al. (2004), Dey (2001)). On comprend donc qu'avec l'avènement des technologies sans fil, la sensibilité au contexte est devenue un caractère incontournable des systèmes qui permettent une utilisation de type *nomade*. Un utilisateur est dit *nomade* s'il peut se connecter au système depuis différents lieux, en utilisant un dispositif d'accès le plus souvent léger, dont il peut changer à tout moment, parfois sans même se déconnecter. La représentation et l'acquisition du contexte sont donc une nécessité pour ces systèmes qualifiés de *pervasifs* ou *ubiquitaires*. La notion de contexte est extensible à volonté. En pratique, elle n'englobe qu'un nombre limité et variable de caractéristiques. Il s'agit principalement des caractéristiques matérielles et logicielles du dispositif d'accès Lemlouma (2004), ou bien uniquement de la localisation de l'utilisateur (Burrell et al. (2002), Rubinsztejn et al. (2004)) ou encore de l'identité de l'utilisateur et de sa localisation (Grudin (2001)). D'autres éléments tels que les bruits environnants, la connexion réseau, la situation sociale de l'utilisateur, etc. influent. Il faut alors considérer le contexte comme l'ensemble des caractéristiques de l'environnement physique ou virtuel qui affecte le comportement d'une application et dont la représentation et l'acquisition sont essentielles à l'adaptation des informations et des services.

Dey (2000) donne une définition générale du contexte qui fait référence : « le contexte est construit à partir de tous les éléments d'information qui peuvent être utilisés pour caractériser la situation d'une entité. Une entité correspond ici à toute personne, tout endroit, ou tout objet (en incluant les utilisateurs et les applications) considéré(e) comme pertinent(e) pour l'interaction entre l'utilisateur et l'application ». L'un des problèmes cruciaux des systèmes sensibles au contexte est celui de la représentation du contexte qui doit aider à en définir les

contours, à conserver les éléments indispensables et à éliminer les informations inutiles (Rey et Coutaz (2004), Brézillon (2002)). Souvent, les contraintes techniques ou financières des capteurs utilisés pour détecter le contexte dictent le choix (Greenberg (2001)) des éléments pertinents à représenter. Bien que réduit, le contexte à représenter n'en est pas moins un espace d'information qui évolue dans le temps (Coutaz et al. (2003), Chaari et al. (2004)). Dans certaines applications, telles que des calculs d'itinéraire ou des guides touristiques (comme les systèmes GUIDE (Cheverest et al. (2002)) ou Campus Aware (Burrell et al. (2002)), la fréquence de la mise à jour et donc de l'activation de la détection du contexte est un paramètre essentiel, fortement lié à la mobilité de l'utilisateur (O'Hare et O'Grady (2002)). Enfin, dans les applications collaboratives, un contexte commun doit pouvoir être partagé (Rey et Coutaz (2004)). Des architectures comme Rover (Banerjee et al. (2002)), MoCA (Rubinsztejn et al. (2004)), ou AWARE (Bardram et Handsen (2004)) permettent la conception de systèmes sensibles au contexte favorisant la collaboration entre utilisateurs équipés de dispositifs mobiles. Elles se présentent comme des API (pour le client et pour le serveur) dotées de plusieurs services pour gérer la collaboration, localiser l'utilisateur, détecter son dispositif, et en fonction de ces informations contextuelles, filtrer l'information et l'adapter à la situation. Les travaux sur l'adaptation dans les systèmes sensibles au contexte se concentrent sur la présentation d'un contenu informationnel sur des dispositifs mobiles (Schilit et al. (2002), Lemlouma et Layaïda (2004)).

2.2 Représentations

Une représentation du contexte en machine doit permettre d'effectuer des raisonnements en vue d'une adaptation. Or, s'il est clair qu'une modélisation exhaustive du contexte est illusoire (Grudin (2001)) en raison de son caractère évolutif, un modèle du contexte doit en contrepartie ne pas être figé. En réalité, les études menées sur les architectures des systèmes sensibles (Mostéfaoui et al. (2004), Chaari et al. (2004)) montrent que la plupart des constructions sont *ad hoc*, complexes, difficiles à modifier et à réutiliser. Par ailleurs, ces systèmes mélangent souvent le code de traitement du contexte avec le code propre à l'application, ce qui augmente considérablement leur complexité. Or, une séparation entre la logique de l'application et celle propre à la gestion du contexte facilite la conception et permet la réutilisation et l'évolution. A ce jour, différents formalismes ont été mis à contribution pour la représentation du contexte : les paires attribut/valeur (utilisé dans les systèmes de Schilit et Teimer (1994) et dans le *Context Toolkit* de Dey (2000, 2001), XML et ses dérivés RDF et CC/PP qui offrent des possibilités de traitement (via XSLT, XQUERY, etc.) et de transmission intéressantes (voir Lemlouma (2004)), les ontologies (définies pour représenter le contexte d'utilisation par Bucur et al. (2005) ou encore par Alarcón et al. (2004) et Leiva-Lobos et Covarrubias (2002) dans le cadre particulier du travail coopératif et du support à la conscience de groupe), les graphes contextuels, proposés par Brézillon (2002) et Mostéfaoui et al. (2004), et les objets qui permettent la représentation du contexte en termes de classes et d'objets, et des associations mettant en relation ces éléments (voir Henricksen et al. (2002) ou Bardram (2005)).

2.3 Acquisition

Outre leur importance, les moyens disponibles pour leur acquisition déterminent les éléments constitutifs du contexte. Le temps et le coût du processus d'acquisition sont décisifs

dans la conception d'un système sensible au contexte à la fois économiquement viable, réactif et interactif. Ainsi, certains aspects du contexte, comme le moment de l'utilisation et la localisation de l'utilisateur, sont plus faciles et moins onéreux à détecter que d'autres, comme l'activité de l'utilisateur. C'est pourquoi la plupart de ces systèmes se contentent de ne traiter qu'une portion du contexte (Burrell et al. (2002)). Vis-à-vis de l'acquisition, on distingue l'information contextuelle (Mostéfaoui et al. (2004)) selon qu'elle peut être détectée, dérivée ou explicite. L'information *détectée* provient de capteurs physiques ou logiciels (température, niveau sonore, pression, altitude, lumière, etc.). La localisation de l'utilisateur entre dans cette catégorie. Elle peut être détectée par GPS en extérieur et par diverses techniques d'approximation en intérieur (par exemple, par croisement de signaux perçus par des bornes Wifi et d'étiquettes électroniques (Anne et al. (2005)). L'information *dérivée* est obtenue lors de l'exécution (comme la date et l'heure). L'information *explicite* est fournie par l'utilisateur. Quel que soit son type, l'acquisition de l'information contextuelle est un processus exposé aux erreurs. L'erreur peut provenir de la panne du capteur, d'un manque de précision, d'un problème de transmission, ou du processus d'interprétation de la donnée brute fournie par un capteur physique vers une donnée raffinée utilisée par le système. Ces erreurs doivent être corrigées par le système ou prises en compte en associant à chaque capteur, voire à chaque donnée reçue, une mesure de fiabilité. Hormis les erreurs, trois difficultés supplémentaires, soulignées par Dey (2000), s'ajoutent à l'acquisition de contexte : *i)* le contexte est dynamique : les changements dans l'environnement doivent être détectés en temps réel et les applications doivent s'adapter à ces changements continus ; *ii)* le contexte est capturé à partir de multiples sources, souvent hétérogènes et reparties ; *iii)* le contexte est obtenu à travers la manipulation de périphériques non conventionnels (autres que la souris ou le clavier). Si aujourd'hui encore la manipulation des capteurs n'est pas aussi bien maîtrisée que celles des périphériques traditionnels, on peut espérer qu'avec l'évolution des standards et techniques tels que Jini, UPnP, et OSGi, capables de gérer la communication et l'interconnexion de dispositifs dans un réseau, la programmation de l'acquisition de données à partir de capteurs sera rapidement chose plus aisée.

Dans l'attente, et afin d'aider le processus d'acquisition, certaines infrastructures comme le Context Toolkit (Dey (2000, 2001)) ou les *contexteurs* (Rey et Coutaz (2004)), ont été proposées. Leur but est d'organiser le processus d'acquisition du contexte indépendamment de l'application. Elles se présentent sous la forme d'un ensemble de composants qui peuvent être assemblés et dont les plus basiques encapsulent le code logiciel nécessaire à la collecte des données auprès des capteurs. Ainsi, grâce à ces infrastructures, il est possible de changer les technologies utilisées pour l'acquisition d'un élément du contexte ou d'ajouter un nouveau composant dédié au traitement d'un nouveau capteur et/ou d'une nouvelle donnée. Si de nombreux verrous scientifiques et technologiques existent encore, l'utilisation combinée d'un système de représentation du contexte et d'une telle infrastructure pour l'acquisition est un premier pas vers la conception de systèmes sensibles au contexte réutilisables et évolutifs.

3 Adaptation des SI collaboratifs

Les travaux décrits ici se situent dans le domaine de l'informatique sensible au contexte. Leur objectif est de faciliter un travail de groupe effectué, via un Système d'Information collaboratif, par des utilisateurs supposés nomades et équipés de dispositifs d'accès mobiles

tels que des PDA, des téléphones cellulaires, etc. L'approche choisie consiste à adapter prioritairement l'information délivrée par de tels SI collaboratifs et, plus particulièrement, l'information dite de *conscience de groupe*. La conscience de groupe regroupe toute connaissance qu'un utilisateur impliqué dans une tâche collective a des activités des autres membres du groupe et de ses propres activités (Dourish et Bellotti (1992), Schmidt (2002)). Cette information sélective doit renforcer la coopération entre les membres du groupe. Les informations de conscience de groupe forment donc un *contexte partagé* de descriptions d'activités individuelles.

L'adaptation mise en œuvre dans ces travaux repose sur une représentation par objets à la fois du contexte d'utilisation et des préférences des utilisateurs. Le modèle de contexte proposé (voir Kirsch-Pinheiro et al. (2004)) comporte – comme dans la plupart des approches classiques – une partie dédiée aux aspects physiques du contexte (localisation, dispositif, application, etc.), et une partie plus originale dédiée à la description ses aspects collaboratifs (notions d'*activité*, de *processus*, de *groupe*, de *rôle*, etc.). Implémenté en AROM, un système de représentation de connaissances par objets (Page et al. (2000), ce modèle est utilisé pour représenter le *contexte courant de l'utilisateur*. Les préférences de l'utilisateur sont prises en compte à travers un ensemble de *profils prédéfinis*, stratifié en différents niveaux de détails par un modèle d'accès progressif présenté dans un de nos premiers travaux sur l'adaptation (voir Villanova-Oliver (2002)). Le mécanisme d'adaptation s'appuie sur un *filtrage* basé sur diverses opérations de comparaison entre des instances AROM. Le filtrage consiste à analyser le contexte courant de l'utilisateur et à sélectionner, parmi les profils prédéfinis, celui qui correspond le mieux à la situation décrite. Puis, l'information de conscience de groupe associée aux profils sélectionnés est délivrée à l'utilisateur, selon l'organisation en niveaux de détail décrite par un modèle d'accès progressif. Cette approche a été implémentée, testée et validée dans une plate-forme pour le support à la conscience de groupe, appelée BW-M (pour *Big Watcher-Mobile*, voir les détails dans Kirsch-Pinheiro (2006)).

3.1 Un modèle à objets du contexte

Comme indiqué, le modèle du contexte proposé (voir Figure 1) intègre des éléments du *contexte physique* et du *contexte collaboratif*. Dans ce modèle, la notion même de *contexte* est représentée par la classe `Description_de_Contexte`. Un objet de cette classe est composé d'objets appartenant aux sous-classes de la classe `Elément_de_Contexte`. Ces sous-classes sont elles-mêmes reliées par des associations[1]. Ainsi, chaque élément du contexte n'est pas isolé mais appartient à un ensemble plus complexe décrivant la situation courante. Par exemple, un membre (classe `Membre`) appartient (association `Apppartient`) à un groupe (classe `Groupe`) à travers les rôles (classe `Rôle`) qu'il tient dans ce groupe. Chaque groupe définit un processus (classe `Processus`) qui est composé (association `Composition`) d'un ensemble d'activités (classe `Activité`), elles-mêmes décomposables en sous-activités, etc.

Ce modèle a été implémenté sous la forme d'une base de connaissances AROM et recouvre trois types de connaissances. Tout d'abord, les objets et les tuples qui instancient ces classes et ces associations décrivent les éléments du système collaboratif et de l'environnement de travail (les groupes, les activités, etc.). Ensuite, cette base stocke les descriptions des profils prédéfinis (voir section suivante) qui correspondent à des contextes

[1] Présentées dans Kirsch-Pinheiro et al. (2004) et Kirsch-Pinheiro (2006).

potentiels et attendus, associés à différents types d'utilisateurs. Enfin, les objets de la classe `Description_de_Contexte` représentent le contexte courant des utilisateurs actifs dans le SI collaboratif. Ces objets sont une connaissance qui est créée dynamiquement et mise à jour par le système durant chaque session de tout utilisateur répertorié. Cette connaissance peut être supprimée lorsque l'utilisateur n'est plus actif. Au contraire, les autres objets (descriptions de la collaboration et des profils) sont stockés en permanence dans la base de connaissances. Ces derniers ne sont cependant pas figés : ils évoluent au gré du travail collaboratif accompli.

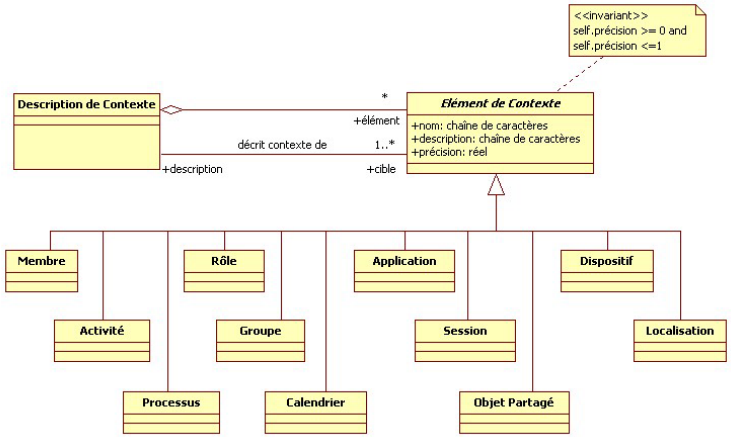


Figure 1. Description UML du contexte.

3.2 Modèles de profil et de conscience de groupe

La conscience de groupe s'organise autour de la création et de la notification d'*événements*. Un événement décrit une information relative à une action qui se déroule (ou qui s'est déroulée, ou encore qui est prévue) dans le processus de coopération (par exemple, la fin d'une activité ou la déconnexion d'un membre de groupe). Au fur et à mesure qu'un groupe progresse dans son travail, que les acteurs de ce groupe interagissent à l'aide du collecticiel, les événements correspondant à ces actions sont générés par le collecticiel et forment l'ensemble des informations de conscience de groupe disponibles.

Le *modèle de conscience de groupe* (ou *modèle de contenu*) est centré sur la classe `Événement` (voir Figure 2) qui comporte un ensemble minimal de variables destiné à décrire une information de conscience de groupe.

Chaque événement est en relation avec un ou plusieurs éléments du modèle de contexte présenté dans la section précédente (voir association `concerne`). De plus, chaque objet de la classe `Événement` est associé, par l'association `est_produit_dans`, à un objet de la classe `Description_de_Contexte` qui représente le contexte dans lequel cet événement est survenu ou va survenir. Par ailleurs, en vue d'un couplage avec le Modèle d'Accès Progressif, la classe `Événement` et ses éventuelles sous-classes sont considérées comme des Entités Masquables potentielles. Des stratifications extensionnelles ou intensionnelles peuvent donc être définies pour ces classes. Par exemple, on peut définir une stratification intensionnelle sur l'ensemble des attributs de la classe `Événement` SI_{int}={{*identifiant,*

description},{intervalle, détails},{médias}}, ou/et une stratification extensionnelle qui organise les objets de cette classe selon la valeur de l'attribut `intervalle` $S2_{ext}$= *{{.intervalle* `during DAY`*}, {intervalle* `during WEEK`*}}*.

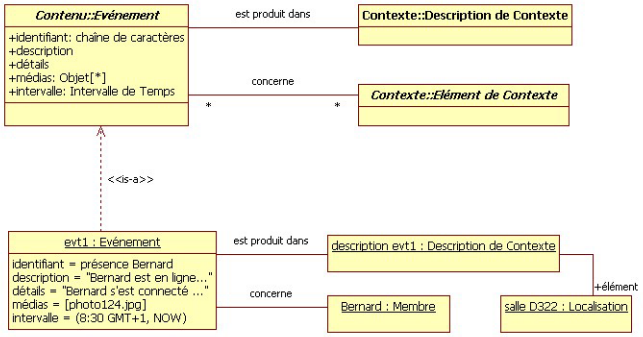


Figure 2. Description UML d'un événement représentant une information de conscience de groupe.

Les préférences de l'utilisateur, quant à elles, sont exprimées à travers la notion de *profil*. Un profil (voir Figure 3) inclut : *i)* la description d'un ou plusieurs contextes potentiels, appelés *contexte(s) d'application,* caractérisant chacun une situation courante d'utilisation ; *ii)* des *règles de filtrage* à appliquer lorsque le contexte courant de l'utilisateur correspond à des contextes potentiels. Ces règles sont formées d'un ensemble de stratifications portant sur des ensembles d'événements. Un profil décrit ainsi les informations pertinentes attendues par ou destinées à un utilisateur dans une situation donnée.

Les profils peuvent être définis par les concepteurs du système collaboratif, ou par les administrateurs des systèmes déployés, ou mieux, par les utilisateurs eux-mêmes. Un concepteur peut créer des profils en fonction des capacités de chaque dispositif d'accès supporté, alors qu'un administrateur peut définir des profils pour des rôles particuliers correspondant à des besoins particuliers en termes d'informations (par exemple, le rôle de coordinateur). Dans le cas d'un utilisateur nomade, une telle fonctionnalité lui permet de définir les situations dans lesquelles il utilise le système, et de définir pour chacune d'elles, ses préférences en matière de contenu. A terme, on peut imaginer que les diverses situations des utilisateurs soient détectées et décrites sous forme de profils générés automatiquement par le système.

3.3 Filtrage par similarité d'objets

Les activités collectives menées au sein d'un travail collaboratif peuvent générer un volume important d'informations de conscience de groupe. Celui-ci peut occasionner une surcharge d'information risquant de perturber les utilisateurs dans l'accomplissement de leurs tâches. De plus, dans le cadre d'une utilisation nomade d'un collecticiel, cette surcharge d'information est difficilement supportée par des dispositifs d'accès mobiles légers aux capacités (écran, mémoire, batterie, etc.) réduites. Le *filtrage* de cette information, pour n'en conserver qu'une partie adaptée et exploitable, est donc nécessaire. Ce filtrage, qui est ici au cœur du processus d'adaptation, est réalisé en deux étapes.

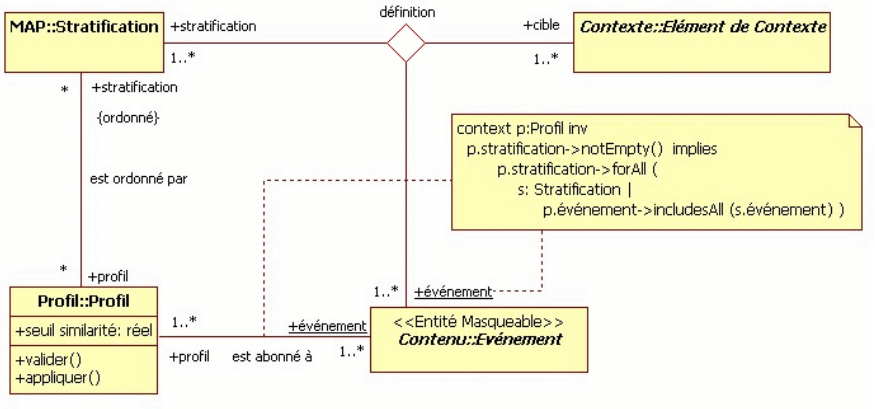


Figure 3. Description UML des liens entre un profil, un ensemble d'événements organisés selon un ensemble de stratifications, et un élément cible du contexte.

La première étape du filtrage opère une *sélection des profils en fonction du contexte courant de l'utilisateur*. Il s'agit de comparer les profils disponibles avec le contexte courant de l'utilisateur. On compare donc deux objets de la classe `Description_ de_Contexte`. Pour chaque profil, on teste si un de ses contextes d'application a un contenu égal ou inclus dans le contenu de la description du contexte courant. Si c'est le cas, le profil correspond à la situation courante, il est donc conservé. Cette comparaison est basée sur une représentation par graphes des objets de la classe `Description_de_Contexte` et des objets de la classe `Eléments_de_Contexte` qui les composent. Dans un tel graphe, les noeuds sont les objets, les arêtes représentent les tuples. Ainsi, un contexte *C* est un sous-contexte d'un contexte *C'* lorsque le graphe associé à *C* est un sous-graphe du graphe associé à *C'*. La relation de sous-graphe est établie à partir de deux opérations *égal* et *contient*[2]. Un nœud *N* est *égal* à un nœud *N'* si l'objet *O* représenté par *N* appartient à la même classe et comporte les mêmes variables que l'objet *O'* représenté par *N'*. Une arête *E* est *égale* à une arête *E'* si les tuples qui les représentent appartiennent à la même association et si ces tuples connectent des objets *égaux*. Un contexte *C contient* un graphe *C'* si, pour chaque nœud *N'* de *C'*, il existe un nœud *N* de *C* tel que *N'* est *égal* à *N* et si, pour chaque arête *E'* de *C'*, il existe une arête *E* de *C* telle que *E'*est *égale* à *E*.

La seconde étape applique *les stratifications fournies par les profils sélectionnés*. Ces stratifications filtrent et organisent les ensembles d'événements. L'algorithme qui régit cette étape : *i)* classe les profils par ordre de priorité ; *ii)* sélectionne parmi les événements auxquels est abonné le profil ceux qui n'ont pas été sélectionnés par un profil précédent dans l'ordre de priorité ; *iii)* applique les stratifications (extensionnelles puis intensionnelles) ; *iv)* délivre le contenu des événements, c'est-à-dire une information de conscience de groupe filtrée. La *priorité* d'un profil est établie par une *mesure de similarité* entre le contexte d'application C_p de ce profil et le contexte courant C_u de l'utilisateur. Il s'agit d'une estimation des éléments du graphe associé au contexte courant qui ont des éléments *égaux* dans le graphe associé au contexte d'application du profil. Les profils dont les contextes d'application sont les plus semblables au contexte courant ont une priorité plus forte. La

[2] Version simplifiée des opérateurs de comparaison et de la mesure de similarité entre objets et tuples AROM proposés (voir Kirsch-Pinheiro (2006) pour plus de détail).

mesure de similarité *Sim* est définie par : $Sim(C_u, C_p) = x$, où $x \in [0, 1]$, tel que x = 1 si chaque élément de C_u a un élément *égal* dans C_p, x = | *X* | / | C_u | sinon, où *X*={ x | x est *égal* à y, x ∈ C_u, y ∈ C_p }.

4 Conclusion

Les travaux présentés ici tentent d'apporter une réponse au problème de l'adaptation posé dans les Systèmes d'Information sur le Web (SIW) dont l'usage est collaboratif. Au-delà de cette spécificité, quel que soit le type de SI abordé, une constante est que l'adaptation repose avant tout sur différents modèles dévolus à la représentation d'un ensemble d'informations ciblées qui seront exploitées par les mécanismes d'adaptation intégrés dans le système. Parmi ces modèles, on trouve le modèle des données du domaine d'application, le modèle des services proposés, le modèle des utilisateurs (et plus généralement, le modèle du contexte d'utilisation), le modèle de la présentation des informations, etc. Nous avons proposé un mécanisme de filtrage de l'information qui tient compte à la fois du contexte de l'utilisateur et de ses préférences dans ce contexte précis. La notion de contexte est représentée par un modèle à objets qui intègre autant le contexte physique de l'utilisateur (sa localisation, le dispositif utilisé…) que le contexte coopératif dans lequel il évolue (notions de groupe, de rôle, d'activité…). Les préférences de l'utilisateur sont représentées par des profils qui permettent de délivrer à l'utilisateur des informations organisées en plusieurs niveaux de détails. Cette proposition a été implémentée au sein d'un canevas nommé BW-M (Kirsch-Pinheiro (2006)).

5 Bibliographie

Anne M., Crowley J.L., Devin V., and Privat G. (2005). *Localisation intra-bâtiment multi-technologies : RFID, Wifi et vision*, UbiMob'05, pp. 29-35, Grenoble, France.

Alarcón R., Collazos C., and Guerrero L.A. (2004), *Distributed shared contexts*, MATA 2004, LNCS 3284, pp. 27-36, Florianópolis, Brazil.

Banerjee S., Agarwal S., Kamel K., Kochut A., Kommareddly C., Nadeem T., Thakkar P., Trinh B., Yossef A., Larson R.L., Shankar A.U., and Agrawala A. (2002). *Rover: scalable location-aware computing*, IEEE Computer, 35(10), pp. 46-53.

Bardram J.E. (2005). *The Java Context Awareness Framework (JCAF) – a service infrastructure and programming framework for context-aware application*, Pervasive'2005, pp. 98-115, Munich, Germany.

Bucur O., Beaume P., and Boissier O. (2005). *Définition et représentation du contexte pour des agents sensibles au contexte*, UbiMob'05, , pp. 13-16, Grenoble, France.

Burrell J., Gray G.K., Kubo K., and Farina N. (2002). *Context-aware computing: a text case*, 4th International Conference on Ubiquitous Computing, LNCS 2498, pp. 1-15.

Bardram J.E. and Hansen T.R. (2004). *The AWARE architecture: supporting context-mediated social awareness in mobile cooperation*, CSCW'04, pp.192-201, Chicago, USA.

Brézillon P. (2002). *Expliciter le contexte dans les objets communicants*, Objets Communicants, pp. 293-303, Hermes Science Publications, Paris.

Brusilovsky P. (1998). *Methods and Techniques of Adaptive Hypermedia*, Adaptive Hypertext and Hypermedia, Kluwer Academic Publishers, pp. 1-43.

Coutaz J., Crowley J.L., Dobson S., and Garlan D. (2003). *Context is the key*, Communication of the ACM, 48(3), pp. 49-53, ACM Press.

Chaari T., Laforest F., and Celentano A. (2004), *Design of context-aware applications based on web services*, Technical Report RR-2004-033, LIRIS, Lyon.

Cheverest K., Mitchell K., and Davies N. (2002). *The role of adaptive hypermedia in a context-aware tourist guide*, Communication of ACM, 45(5), pp. 47-51.

Dourish P. and Bellotti V. (1992). *Awareness and Coordination in Shared Workspaces*, ACM Conference on Computer-Supported Cooperative Work, ACM Press, pp 107-114.

Dey A.K. (2000). *Providing Architectural Support for Building Context-Aware Applications*, PhD Thesis, Georgia Institute of Technology.

Dey A.K. (2001). *Understanding and using context*. Personal and Ubiquitous Computing, 5(1), pp. 4-7.

Frasincar F. and Houben G.-J. (2002). *Hypermedia Presentation Adaptation on the Semantic Web*, AH 2002, LNCS 2347, pp. 133-142.Malaga, Spain.

Greenberg S. (2001). *Context as a dynamic construct*, Human-Computing Interaction, 16(2-4), pp. 257-268.

Grudin J. (2001). *Desituating action: digital representation of context, Human-Computing Interaction*, 16(2-4), pp. 269-286.

Henricksen K., Indulska J., and Rakotonirainy A. (2002). *Modeling context information in pervasive computing systems*, Pervasive'2002, pp. 167-180, Zürich, Switzerland.

Kirsch-Pinheiro M., Gensel J., and Martin H. (2004). *Representing Context for an Adaptative Awareness Mechanism*. CRIWG'04, LNCS 3198. Springer, pp 339-348.

Kirsch-Pinheiro M. (2006). *Adaptation contextuelle et personnalisée de l'information de conscience de groupe au sein des Sytèmes d'Information coopératifs*. Thèse de doctorat, Université Joseph Fourier, Grenoble, France.

Koch. N. (2000). *Software Engineering for Adaptative Hypermedia Systems – Reference Model, Modelling Techniques and Development Process*, Ph.D Thesis, Fakultät der Mathematik und Informatik, Ludwig-Maximilians-Universität München.

Leiva-Lobos E.P. and Covarrubias E. (2002), *The 3-ontology: a framework to place cooperative awareness*, CRIWG 2002, LNCS 2440, pp. 189-199.

Lemlouma T. (2004), *Architecture de négociation et d'adaptation de Services Multimédia dans des Environnements Hétérogènes,* Thèse de Doctorat, INPG, Grenoble, France.

Lemlouma T. and Layaïda N. (2004). *Context-Aware Adaptation for Mobile Devices*, IEEE International Conference on Mobile Data Management, pp. 106-111.

Mostéfaoui K., Pasquier-Rocha J., and Brézillon P. (2004). *Context-aware computing: a guide for the pervasive computing community*, IPCS'04, IEEE Computer, pp. 39-48.

O'Hare G. and O'Grady M. (2002). *Addressing mobile HCI needs through agents*, Mobile HCI 2002, LNCS 2411, pp. 311-314.

Page M., Gensel J., Capponi C., Bruley C., Genoud P. and Ziébelin D. (2000). *Représentation de connaissances au moyen de classes et d'associations : le système AROM*, LMO 2000, pp 91-106, Mont Saint Hilaire, Québec, Canada, January 26-28.

Paterno P. and Mancini C. (1999). *Designing Web User Interfaces for Museum Applications to Support different Types of Users,* International Conference about Museums and the Web, pp.75-86, LA, USA.

Raad H. and Causse B. (2002). *Modelling of an Adaptive Hypermedia System Based on Active Rules*, ITS 2002, LNCS 2363, pp. 149-157, Biarritz, France.

Rey G. and Coutaz J. (2004). *Le contexteur : capture et distribution dynamique d'information contextuelle*, UbiMob'04, pp. 131-138, Nice, France.

Rubinsztejn H.K., Endler M., Sacramento V., Gonçalvez K., and Nascimento F. (2004). *Support for context-aware collaboration*, MATA 2004, LNCS 328, pp.37-47, Florianópolis, Brasil.

Schmidt K. (2002). *The problem with 'awareness': introductory remarks on 'Awareness in CSCW'*., Computer Supported Cooperative Work, 11(3-4), pp 285-298.

Schilit B.N., Hilbert D.M., and Trevor J. (2002). *Context-aware communication*, IEEE Wireless Communications, 9(5), pp. 37-45.

Schilit B.N. and Theimer M.M. (1994). *Disseminating active map information to mobile hosts*, IEEE Network, 8(5), , pp. 22-32.

Villanova-Oliver M. (2002). *Adaptabilité dans les systèmes d'Information sur le Web : Modélisation et mise en œuvre de l'accès progressif*, Thèse de Doctorat, Institut National Polytechnique de Grenoble.

Modèle contextuel pour la personnalisation

Zeina Jrad*, Abdouroihamane Anli*
Marie-Aude Aufaure*,**

*INRIA Paris-Rocquencourt, Domaine de voluceau
Rocquencourt B.P. 105, 78153 LeChesnay cedex France
http://www.inria.fr
{Zeina.Jrad, Abdouroihamane.Anli}@inria.fr

** Supélec – Département informatique
Plateau du Moulon
91192 Gif sur Yvette cedex
http://www.supelec.fr
Marie-Aude.Aufaure@supelec.fr

Résumé. Les travaux sur la personnalisation ont fourni récemment des résultats importants concernant la recherche d'information sur le Web et les systèmes de recommandation. L'objectif principal d'un système de personnalisation est de fournir des informations en prenant en compte les préférences des utilisateurs et les informations contextuelles. Cet article présente un modèle qui représente l'utilisateur et son contexte de navigation dans un portail de tourisme[1]. Notre proposition se situe dans le cadre du projet Eiffel, qui vise à développer un moteur de recherche sémantique dédié au tourisme. Le papier présente également la façon dont le modèle est alimenté par extraction de données en utilisant les techniques du Web Usage Mining.

1 Introduction

Le Web actuel peut être vu comme une immense bibliothèque de documents aux formats et contenus très hétérogènes. Compte tenu de la croissance continue de ces documents, il devient de plus en plus difficile de trouver les ressources pertinentes qui répondent à une requête posée par un utilisateur. De plus cette bibliothèque évolue de manière souvent imprévisible, soit par ajout/suppression de documents, soit par modification du contenu des documents existants.

Les moteurs de recherche disponibles sur le Net renvoient habituellement plus de 1.500 résultats par question Zemirli et al. (2005). Pourtant parmi les vingt principaux résultats,

[1] Ce travail a été financé en partie par l'Agence Nationale de la Recherche, dans le cadre du projet RNTL Eiffel (web sémantique et e-tourisme).

seulement la moitié d'entre eux sont susceptibles d'être appropriés aux besoins de l'utilisateur. Le problème n'est plus tant la disponibilité de l'information mais sa pertinence relativement aux besoins précis d'un utilisateur, dérivés à partir de ses préférences, son contexte et les représentations qu'il perçoit.

Les documents présents sur le Web ne sont pas des documents statiques consultés passivement, mais des documents souvent générés à la demande (documents virtuels), et dans lesquels la consultation implique une participation active de l'utilisateur. Ce dernier point rend donc importante la notion de personnalisation de ces documents virtuels, afin de faciliter leur consultation en tirant avantage de leur grande "contextualité".

Un système capable de fournir une interaction personnalisée nécessite un modèle utilisateur Garlatti et Prié (2003).Une interaction personnalisée nécessite plusieurs étapes parmi lesquelles Razmerita (2005a) :

- la définition du modèle utilisateur
- l'acquisition des données utilisateurs
- le raisonnement et les inférences
- la génération de services personnalisés

Les mécanismes de personnalisation peuvent consister en des techniques adaptatives variées ou être basées sur les interactions des agents. Un système adaptatif, flexible permet l'adaptation de sa fonctionnalité et son contenu hypermédia selon les besoins et caractéristiques des utilisateurs. Au niveau de l'interface utilisateur les techniques d'adaptation peuvent être classifiées, selon Kobsa et al. (2000), en trois catégories : adaptation de la structure, adaptation du contenu, adaptation de la modalité et de la présentation. Une description plus détaillée des différentes techniques d'adaptations dans le contexte des systèmes de gestion de connaissance peut être trouvé en Razmerita (2005).

Nous nous intéressons dans nos travaux à un moteur de recherche pour le tourisme dans le cadre du projet Eiffel (2005) qui fournit à l'utilisateur un outil d'aide à la recherche d'information et à la navigation Jrad et Aufaure (2007). Ce moteur de recherche et de navigation s'appuie sur le/les web services de requête sémantique aussi bien que sur le profil et les préférences des utilisateurs. Notre travail vise donc à proposer à l'utilisateur, dans le cadre du projet Eiffel, une recommandation d'items basés sur le profil. Cette recommandation a pour principale fonction d'agréger des contenus émanant de différentes sources du portail et d'en personnaliser la présentation pour chaque utilisateur selon son profil. Elle permet de restreindre l'affichage des résultats de la recherche à une sélection qui est en accord avec le profil de l'utilisateur. Des suggestions de destinations sont proposées à l'utilisateur en fonction de ses intérêts et de son profil. Pour chaque destination, des objets associés à cette destination sont affichés afin de montrer à l'utilisateur que la destination proposée correspond bien à ses centres d'intérêts et afin qu'il ait accès aux activités ou aux sites reliés à cette destination de manière directe.

2 Conception d'un modèle "*context-aware"*

2.1 Modélisation des navigations

Il existe principalement trois types de représentation de profil :

- *ensembliste* : le profil y est généralement formalisé comme des vecteurs de termes pondérés Budzik et Hammond (2000), Dumais et al. (2003) ou classes de vecteurs Mc Gowan (2003).
- *multidimensionnelle* : le profil est structuré selon un ensemble de dimensions, représentées selon divers formalismes Bouzeghoub et Kostadinov (2004), Zemirli et al. (2005), Anli (2006).
- *Sémantique* : la représentation du profil met en évidence les relations sémantiques entre informations le contenant. La représentation est essentiellement basée sur l'utilisation d'ontologies ou réseaux sémantiques probabilistes Lin et al. (2005).

La représentation des connaissances à base d'ontologie a récemment retenu l'attention comme un domaine de recherche très prometteur pour le développement d'une nouvelle génération de systèmes d'information et pour le développement du Web. L'ontologie représente aussi une notion clé pour le développement du Web sémantique.

L'ontologie permet la représentation des connaissances à base d'une conceptualisation. L'ontologie peut être définie aussi comme un vocabulaire conceptuel consensuel. Sowa (2000) définit une ontologie comme un domaine qui a comme objet l'étude des catégories qui existent ou qui pourraient exister dans un certain domaine. Les langages d'ontologie orientés Web actuels comme: OWL, SKOS, DAML+OIL, etc. repoussent et étendent les standards Web proposés par W3C, notamment XML, RDF/RDFS.

L'idée de créer des structures de connaissances réutilisables contenant des caractéristiques et préférences des utilisateurs est envisagée pour permettre l'interaction ubiquitaire, contextualisée pour le bénéfice des utilisateurs.

La modélisation utilisateur est un élément clé pour l'interaction personnalisée. La modélisation utilisateur à base d'ontologie nécessite une structure référentielle classique statique et une partie adaptative qui doit évoluer à partir de progrès de l'apprenant conformément à son but, domaine d'intérêt, etc.

Gavrilova et al. (2006) proposent une méta-ontologie du domaine de modélisation de l'utilisateur. Cette approche structurée basée sur la méthodologie de l'ingénierie cognitive peut faciliter les travaux de recherche. Elle comporte l'ontologie ou la structure conceptuelle et l'amélioration de domaine comme guide pour les travaux de recherche significatifs. Cette ontologie est censée structurer l'état de l'art dans le domaine et sert comme un point de référence central et d'outil de guidage pour les travaux de recherche sur la modélisation.

Les systèmes utilisateur adaptatifs sont au coeur de la conception des initiatives d'intelligence ambiante. Évidemment, les modèles d'utilisateur sont les "ingrédients" nécessaires pour de tels systèmes. Kikiras et al. (2006) présentent un modèle d'utilisateur pour des systèmes de navigation (principalement pour les piétons), qui est basé sur des théories de navigation humaine. Ils présentent ce modèle par une ontologie de Web sémantiques et montrent comment elle peut être incorporée dans un système de navigation d'intérieur appelé Onto-Nav, qui permet la sélection personnalisée de chemin.

Les systèmes et les applications stockent des préférences et des informations sur des utilisateurs afin de fournir un accès personnalisé. Cependant, ces systèmes stockent des profils utilisateur dans des formats possessifs. L'échange ou la réutilisation d'information n'est pas possible et l'information est reproduite. Mehta et al. (2005) proposent l'utilisation d'un modèle utilisateur contextuel basé sur une ontologie comme une base pour l'échange de profils utilisateur entre différents systèmes.

2.2 Détermination des composants de notre modèle

Dans notre travail, le contexte de l'utilisateur est décrit par un ensemble de facettes. L'utilisateur interagit avec le système dans différents rôles et est impliqué dans des taches parallèles dont chacune est associée à un sous-ensemble spécifique de facettes du contexte de l'utilisateur. Pour refléter cette structure, le contexte de l'utilisateur est divisé en multiples contextes fonctionnels (figure 1) regroupant les facettes du contexte utilisateur reliés au même rôle ou tâche.

Selon la définition présentée ci-dessus, nous avons établi un modèle regroupant les catégories/composants suivants (figure 1) : Contexte_Utilisateur = Profil_Utilisateur U Ancien_Parcours U Parcours_Courant U Contextes_Off U Contexte_Materiel U Contexte_Spatial.

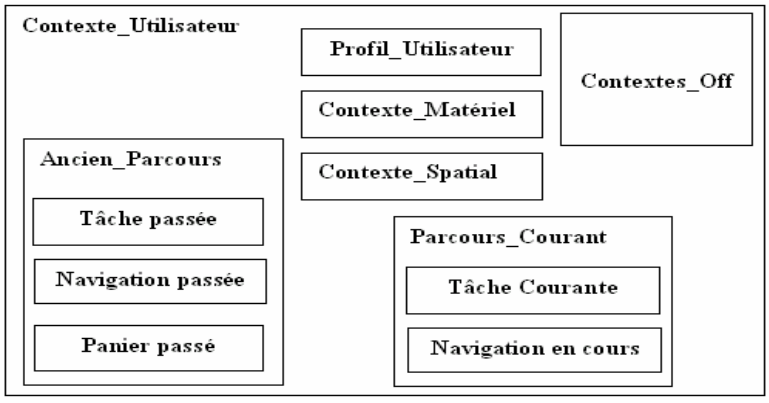


FIG 1 – *Définition du Context_Utilisateur*

Profil_Utilisateur = informations rémanentes indiquant toutes connaissances (points d'intérêts, capacités budgétaires, …) que le système a pu avoir sur l'utilisateur directement (formulaire/inscription) ou indirectement (déduction/requête). Ceci représente les tendances de l'utilisateur au moment de la connexion qui restent vrai tant que non modifiées par lui ou par le système et inclut :

- Les données personnelles : elles sont la partie statique du profil. Elles comprennent l'identité de l'utilisateur (nom, prénom, numéro de sécurité sociale, etc.), des données démographiques (age, genre, adresse, situation familiale, nombre d'enfants, etc.), les contacts personnels et professionnels de l'utilisateur et d'autres informations comme le numéro de la carte bancaire ou de la carte Vitale. Les données personnelles sont relativement stables dans le temps et ne demandent pas de mise à jour automatique par le gestionnaire de profils. Ces données peuvent ne pas jouer un rôle dans le processus de recherche d'information, mais servent comme monnaie d'échange contre les services de personnalisation ou les services d'accès. Elles sont récupérées en demandant aux utilisateurs de fournir des formulaires de données per-

sonnelles qui sont ensuite utilisées pour faire des statistiques pour mieux cibler les clients.

- Le centre d'intérêt : Le centre d'intérêt exprime le domaine qui intéresse l'utilisateur ou son périmètre d'exploration. Le centre d'intérêt est vu comme une présélection virtuelle qui réduit la masse d'informations à prendre en compte. Par conséquent toute requête émise par l'utilisateur sera enrichie avec les mots clés ou les prédicats des requêtes définissant le centre d'intérêt.

Ancien_Parcours = Tâche passée + Navigation passée + Panier de sélection

La Tâche passée indique la raison pour laquelle l'utilisateur est rentré sur le site lors de son dernier login c'est à dire le but de la navigation. Elle peut être défini par un ensemble de mots clés ou un ensemble d'expressions logiques (requêtes) dans une session de navigation précise. La Navigation passée regroupe ce qui a été effectuée pour accomplir la Tâches passées et le Panier de sélection représente les objets touristiques choisis et les réservations effectuées par l'internaute pour construire son voyage ou ses vacances. L'ancien parcours constitue un historique de navigation qui aide à tracer le parcours complet d'un utilisateur qui s'est connecté sur le portail de tourisme pour planifier un voyage.

Parcours_Courant = Tâche courante + début de Navigation en cours

La Tâche courante décrit ce que l'utilisateur est en train de chercher sur le site en temps réel alors que la navigation en cours indique l'historique de ses actions depuis le début de la session en cours.

Contextes_Off = Contextes_Utilisateurs des autres navigations par d'autres utilisateurs.

Contexte_Materiel = Type de terminal utilisé par l'utilisateur durant la navigation. Ceci peut être un portable, un PC, un PDA, etc. Il est nécessaire pour l'adaptation de l'interface graphique selon les dimensions du terminal d'accès.

2.3 Représentation des connaissances à base d'ontologie

Le modèle utilisateur est le résultat du processus de modélisation de l'utilisateur. D'habitude, le modèle contient des informations sur les buts, les besoins, les préférences ou les intentions des utilisateurs. Les modèles utilisateur plus avancés peuvent contenir des informations liées à l'état psychique, émotionnel, physique, etc.

Notre modèle, représenté graphiquement en Figure 2, a été implémenté en utilisant IModeler et KAON, une boite à outils pour la gestion d'ontologie et une A.P.I. pour le développement d'applications basées sur les ontologies Maedche et al. (2003).

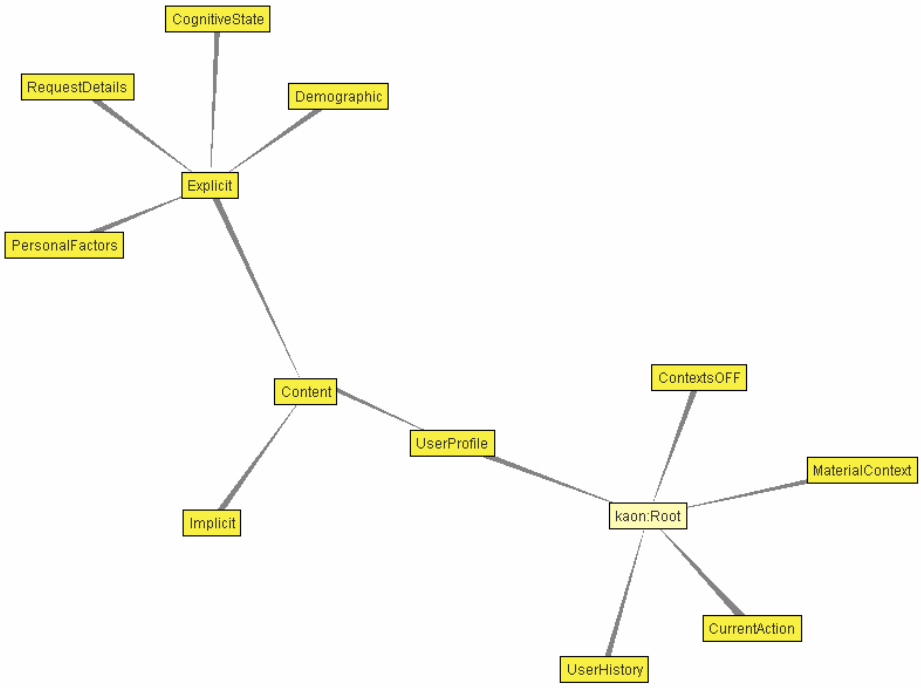


Fig. 2 – Une partie du model contextuel représenté dans OI Modeler

3 Web Usage Mining et Recherche d'informations touristiques personnalisées

Le Web Usage Mining consiste à améliorer la navigation sur le Web en utilisant des informations sur la façon dont les utilisateurs naviguent. Cette amélioration peut concerner les performances des serveurs, la structure du site, la personnalisation, l'aide à la navigation, les recommandations de produits ou de services, etc.

L'analyse des fichiers logs permet de comprendre le comportement des internautes sur le site web. Le but est de calculer des statistiques sur les usages du navigateur, afin de créer des profils utilisateurs et d'optimiser les interfaces de formulation de requêtes, de navigation et d'affichage des résultats. Ces profils correspondent à des modèles comportementaux et permettront la personnalisation de l'interface du moteur de recherche, aussi bien pour la formulation des requêtes que pour l'affichage des résultats.

Le WUM nous permet d'améliorer le portail touristique du point de vue navigation et recherche de la façon suivante :

- Présenter à l'utilisateur une sorte d'assistance lors de sa connexion sur le site de telle façon à ce que le contenu du site soit en lien avec le profil de l'utilisateur. Par exemple, proposer à un utilisateur intéressé par la culture des offres intéressantes sur des musées dans la région même si sa requête n'inclut pas explicitement le mot musée.

- Proposer à un utilisateur dès sa connexion sur le site s'il souhaite effectuer une recherche personnalisée. Si c'est le cas, on lui propose des recommandations en rapport avec ses centres d'intérêt.
- Aider l'utilisateur à formuler sa requête en mettant plus en avant les catégories dans la taxonomie proposée qui risquent de l'intéresser. Le reste de la facette, menu ou icône peut être affiché en petite taille ou en italique en bas de la page.
- Proposer dès la page d'accueil des séjours complets selon les différents types de profils affichés et présentés à l'utilisateur. Par exemple, le Profil A concerne une catégorie d'utilisateur véhiculé qui voyage en famille accompagné d'enfants à budget limité pour une période importante (2 à 3 semaines). Ou bien, le Séjour X contient une proposition de programme complet sur 2 ou 3 semaines dans la région avec un calendrier des visites et activités à effectuer chaque jour, les routes à prendre, le coût de chaque visite....le tout en parfaite harmonie avec les caractéristiques du profil A.
- Comprendre la perception du site : classer le contenu du site en fonction de l'usage :
  - découvrir les rapprochements réalisés par les utilisateurs entre les documents du site
  - surveiller le contenu
  - améliorer l'organisation
  - visualiser la classification pour faciliter sa compréhension par les administrateurs du site
  - visualiser la façon dont le site est perçu en dehors de toute classification

4 Utilisation du Web Usage Mining pour l'alimentation du modèle contextuel

Un processus WUM comporte trois étapes principales : prétraitement, fouille de données et analyse des motifs extraits. Dans ce papier, nous nous intéressons à l'étape de prétraitement (nettoyage, sélection et transformation des données issues des fichiers logs) pour la préparation des données issues des fichiers logs dans une perspective d'extraction de connaissances utiles au modèle contextuel présenté plus haut. Plusieurs techniques existent pour le prétraitement des logs. L'approche de Tanasa et Trousse (2004) a été utilisée pour aboutir à un premier schéma de données.

Ce schéma est ensuite étendu pour supporter d'autres informations utiles au modèle contextuel. La figure 3 présente le modèle obtenu à l'issu du prétraitement. Les tables dans les cadres grisés correspondent à ceux que nous avons ajoutées dans le modèle initial proposé par Tanasa et Trousse. Les principales extensions portent sur la reconnaissance de la localisation, l'identification des tâches de l'utilisateur et l'extraction des objets touristiques manipulés ou recherchés par les utilisateurs.

Reconnaissance de la localisation : Il s'agit de retrouver le pays, la région ou même la ville où se trouve l'utilisateur interagissant avec le système. La technique utilisée consiste à une mise en correspondance entre les adresse IP et les localisations géographiques. Des outils sont disponibles, par exemple, l'API GeoIP (http://www.maxmind.com) permet de déterminer les localités en fonction des adresses IP. Cette information permet de renseigner les composants *Contexte_Spatial* du modèle contextuel.

Identification des tâches : Une tâche correspond à l'ensemble des mots clés manipulés par l'utilisateur durant sa navigation. Ces mots clés sont récupérés à partir des paramètres des URLs et permettent de renseigner les champs *tâche passée* et *tache courante* du modèle contextuel.

Extraction des objets touristiques : Dans le cadre de notre application, les objets touristiques se présentent sous forme de pages HTML (donc des URLs, plus précisément des fichiers) issues d'une transformation XSL à partir de documents XML. Chaque page HTML contient un identifiant permettant d'effectuer la correspondance avec son XML associé. Ces données stockées dans les fichiers XML permettent de connaître, par exemple, les centres d'intérêts ou les paniers passés des utilisateurs.

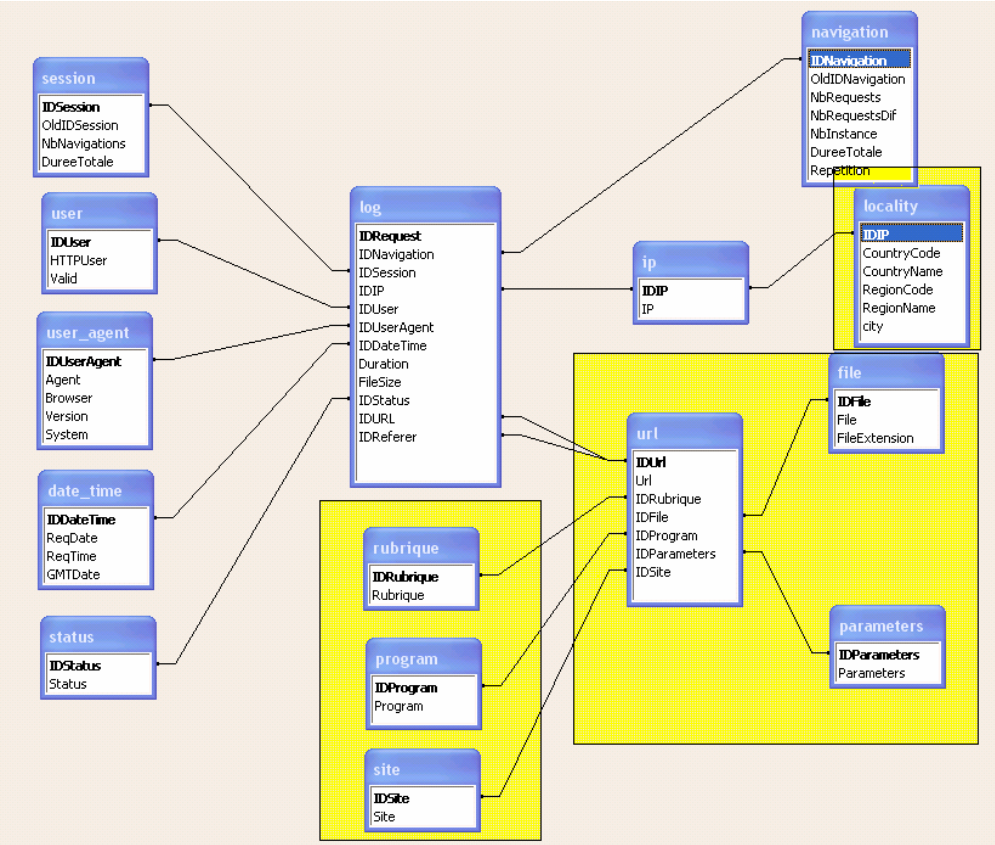


FIG. 3 – Modèle de données.à l'issu de prétraitement [extension du modèle de Tanasa et Trousse (2004)]

5 Conclusion

Dans cet article, nous avons présenté quelques travaux sur la modélisation contextuelle des utilisateurs pour la personnalisation. Nous avons décrit un modèle basé sur le contexte de navigation des utilisateurs. Ce modèle contient plusieurs catégories comme l'historique de navigation, les préférences, la tache courante. Le model étant dynamique et progressivement extensible, nous avons montré dans le papier comment l'enrichir grâce à l'extraction de données par les outils du Web Usage Mining, notamment, pendant la phase de préparation des données. Une perspective à moyen terme de ces travaux concerne l'utilisation des connaissances produites pendant la phase de fouille de données pour affiner le modèle contextuel de l'utilisateur.

Références

Eiffel 2005. Eiffel *Web Sémantique et e-Tourisme*, Projet RNTL (Réseau National de recherche et d'innovation en Technologies Logicielles), INRIA et al., Livrable de projet.

Anli, A. (2006) *Méthodologie de développement des systèmes d'information personnalisés*. Thèse de doctorat, Université de Valenciennes et du Hainaut-Cambrésis.

Bouzeghoub, M. et D. Kostadinov (2004). *Une approche multidimensionnelle pour la personnalisation de l'information*. Rapport PRiSM, Versailles, France.

Budzik, J., K. Hammond (2000). User interactions with every applications as context for just-in-time information access, Proceedings of *the 5th international conference on intelligent user interfaces*, 44-51.

Dumais, S., E. Cuttrel, J. Cadiz, G. Jancke, R. Sarin and D. Robbins (2003). Stuff I've seen : A system for a personal information retrieval and re-use. Proceedings of the *26th ACM SIGIR International Conference on Research and Development,* 72-79.

Garlatti, S. et Y. Prié (2003).Adaptation et personnalisation dans le Web sémantique. Numéro spécial de la *revue Information, interaction, Intelligence*.

Gavrilova, T., P. Brusilovsky, M. Yudelson and S. Puuronen (2006). Creating Ontology for User Modelling research. In: Workshop on *Ubiquitous User Modeling* in conjunction with the *17th European Conference on Artificial Intelligence (ECAI 2006).*

Jrad, Z. and M. A. Aufaure (2007). Personalized Interfaces for a Semantic Web Portal: Tourism Information Search. In the *11th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, SWEA – KES'2007

Kikiras, P., T. Vassileios and H. Stathes (2006). Ontology-Based UserModeling for Pedestrian Navigation Systems. In: Workshop on *Ubiquitous User Modeling* in conjunction with *the 17th European Conference on Artificial Intelligence (ECAI 2006).*

Kobsa, A., J. Koenemann, and W. Pohl (2000). Personalized hypermedia presentation techniques for improving online customer relationships. *The Knowledge Engineering Review* 16, 111-155.

Lin, C., G. Xue, H. Zeng and Y. Yu (2005). Using probabilistic latent semantic analysis for personalised Web search., Proceedings of *the APWeb Conference,* 707-711.

Maedche, A., B. Motik, L. Stojanovic, R. Studer and R. Volz (2003). Ontologies for Enterprise Knowledge Management., *IEEE Intelligent Systems*.

MC Gowan, J. (2003) *A multiple model approach to personnalised information access*. Master thesis in computer science, Faculty of science, University college Dublin.

Mehta, B., C. Niederee, A. Stewart, M. Degemmis, P. Lops, G. Semeraro (2005). Ontologically-Enriched Unified User Modeling for Cross-System Personalization. In: Proceedings of *User Modelling 2005*, published by Springer Berlin/Heidelberg, 119-123.

Razmerita, L. (2005a) User modeling and personalization of the Knowledge Management Systems, book chapter, in *Adaptable and Adaptive Hypermedia*, published by Idea Group Publishing, 225-245.

Razmerita, L. (2005b) Services contextualisés pour utilisateurs et la modélisation des utilisateurs à base d'ontologies : défis et perspectives. Atelier sur *la Modélisation Utilisateur et la Personnalisation d'Interfaces Homme-Machine*, en conjonction avec la conference *Extraction et Gestion des Connaissances*, Paris.

Tanasa, D. and B. Trousse(2004). Advanced Data Preprocessing for Intersites Web Usage Mining. In *IEEE Intelligent Systems* 19(2):59-65.

Zemirli, N., L. Tamine et M. Boughanem (2005). Accès personnalisé à l'information : Proposition d'un profil utilisateur multidimensionnel. *7th ISPS'Algiers*.

Summary

In this paper, we have presented some background knowledge on modeling theory and personalization facilities related to the proposed contextual model. We focuse on how to model the user and his context in an extensible way that can be interpreted and used for personalization. We have also described data extraction in order to extend our model using WUM facilities. This work is actually in progress as a part of the Eiffel project.

Recherche de patrons de navigation pour les systèmes de recommandations

Ali Mroué*, Jean Caussanel*

*Laboratoire des Sciences de L'Information et des Systèmes
LSIS - UMR CNRS 6168
Université Paul Cézanne, Aix-Marseille III
Domaine Universitaire de St Jérôme
Ave. Escadrille Normandie-Niemen
13397 Marseille Cedex, France

{ali.mroue, jean.caussanel}@lsis.org,
http://www.lsis.org/~ ali_mroue.html
http://www.lsis.org/~ jean_caussanel.html

Résumé. Cet article décrit une approche pour rechercher automatiquement les comportements prototypiques dans un ensemble de parcours recueillis pour un site web. Les fichiers de traces de navigations (fichiers access log) sont examinés afin de grouper les utilisateurs qui ont un patron d'accès commun et fréquent. Ces résultats nous permettent de comprendre la manière avec laquelle les internautes évoluent sur un site donné et, à plus long terme de déceler des pratiques générales de navigation.Une application résultante de ce type de recherche se situe dans les systèmes de recommandation où l'on cherche à suggérer des liens aux utilisateurs qui pourraient les intéresser. Après avoir sommairement rappelé les principes des techniques des méthodes existantes dans ce domaine nous décrivons les caractéristiques de la fonction de similarité que nous proposons. Nous présentons ensuite nos résultats obtenus sur un site particulier.

1 Introduction

Le World Wide Web fournit un environnement riche pour la recherche d'information. Jour après jour, l'Internet se développe et la quantité d'information disponible devient si importante que les utilisateurs peuvent facilement se perdre dans cette grande source d'information et ce, malgré l'aide des moteurs de recherche. Le Web Mining est un domaine qui propose des solutions, entres autres, pour l'aide à la recherche. Fondamentalement, ce domaine consiste à utiliser l'ensemble des techniques du Data Mining afin de développer des approches et des outils, permettant d'extraire des informations pertinentes à partir des données du web (documents, traces d'interactions, structure des pages, des liens, etc. . .)

2 Web Usage Mining

Dans cet article nous nous intéresserons au Web Usage Mining, pour résoudre les problèmes évoqués ci-dessus. Le Web Usage Mining est l'analyse du comportement de l'utilisateur à travers sa navigation et notamment l'ensemble des clics effectués sur le site (clickstream). On peut par exemple déterminer les navigations les plus fréquentes afin d'améliorer le site ou le rendre adaptatif (Perkowitz et O., 2000). Le Web Usage Mining exploite des techniques du Data Mining sur les fichiers logs des serveurs, et dans l'utilisation des cookies d'un site donné.

2.1 Principe Général

Le Web Usage Mining se base en général sur un traitement des données en trois phases : prétraitement "preprocessing", découverte de modèle "pattern discovery", analyse des modèles "pattern analysis" (Srivastava et al., 2000).

Prétraitement Cette phase consiste à reconstruire les sessions d'utilisateurs à partir du fichier log. Dans notre travail nous avons épuré les fichiers log en nous basant sur les méthodes proposées par (Cooley et al., 1999), qui consistent en quatre étapes, afin de reconstruire les sessions : Data Cleaning, User Identification, Session Identification, Path completion.
Signalons que l'étape de prétraitement, bien que très technique, est d'une importance majeure dans le processus puisque'elle nettoie les traces d'actions des éléments non demandés par l'internaute mais implicitement chargés par le serveur web. Une erreur dans cette phase falsifie de fait la séquence receuillie.

Découverte des modèles Cette phase consiste à exploiter les données filtrées résultantes de l'étape précédente afin de découvrir des modèles comportementaux qui décrivent les navigations des utilisateurs.

Analyse des modèles L'analyse des modèles est la dernière étape globale du Web Usage Mining. Elle a comme objectif de filtrer les modèles inintéressants de l'ensemble trouvé dans la phase de découverte. Ce filtrage dépend de l'application finale que l'on souhaite faire du Web Usage Mining (adaptation des sites web, système de recommendation, préchargement des pages, etc…).

2.2 Applications

Les travaux sur le Web Usage Mining ont été exploités dans plusieurs types d'applications. Ces applications ont le plus souvent pour objectif d'aider les utilisateurs à s'orienter et trouver plus rapidement l'information cherchée. Dans ce cadre on peut citer par exemple les systèmes de recommandation (Baoyao et al., 2004), qui sont conçus pour aider les utilisateurs à prendre des décisions dans un espace complexe où la quantité d'information à disposition est très importante. Les systèmes de recommandation exploitent un modèle de prédiction pour anticiper sur l'intérêt des utilisateurs et faire ensuite des recommandations. Ils consistent à prédire les prochaines actions de l'utilisateur en se basant sur leurs actions précédentes.

D'autres applications du Web Usage Mining visent le "cache" et le préchargement des pages "caching and prefetching". Beaucoup d'utilisateurs considèrent que la "latence" (Intervalle de temps qui sépare la demande de la réponse) est un problème important du Web. Beaucoup de facteurs contribuent à ce problème de latence : la latence de transmission, consultations de nom de domaine " DNS ", établissement de raccordement de TCP, le délai de début de session dans le serveur Http (Cohen et Haim, 2000). Les techniques Web conventionnelles de "caching" (Aggarwal et al., 1999; Shim et al., 1999) visent à résoudre une partie de ce problème en stockant temporairement, au plus près de l'utilisateur, le contenu récemment accédé par celui-ci, sur le dispositif du client ou sur un serveur proxy. Cependant, "caching" peut ne pas réduire la latence dans le cas oú l'accès se fait sur des contenus très dynamiques et personnalisés. Une approche complémentaire pour réduire la latence consiste à prédire le comportement de l'utilisateur et d'employer cette connaissance pour précharger sa future demande au plus près de l'utilisateur (Sarukkai et R., 2000).

Enfin on peut trouver des applications du Web Usage Mining pour l'amélioration des moteurs de recherche (Baeza-Yates, 2004), et la personnalisation des sites Web (Pitkow et Peter, 1999; Baoyao et al., 2005).

Si dans ces applications l'objectif recherché peut se généraliser comme une prédiction de la future page visitée par l'internaute, les techniques pour déterminer cette page sont assez variées. Nous allons dans le paragraphe ci-dessous présenter les principales techniques du Web Usage Mining.

2.3 Principales Techniques Du Web Usage Mining

Les principales techniques pour déterminer les modèles de navigation des utilisateurs d'un site web sont fondées sur : l'analyse statistique, des règles ad-hoc d'association, des techniques de clustering, des méthodes de classification, ou la recherche des motifs séquentiels. Dans beaucoup de travaux ces techniques sont combinées afin de trouver des modèles de navigations les plus proches des parcours réels des utilisateurs. Dans les paragraphes suivants nous donnerons quelques éléments d'information concernant chacune de ces techniques.

Analyse Statistique La plupart des applications commerciales concernant le Web Usage Mining sont fondées sur les techniques d'analyse statistique. En appliquant différentes méthodes d'analyse statistique (fréquence, médiane, moyen, etc…), on peut extraire des informations, telles que les pages les plus souvent accédées, le temps moyen de visite d'une page ou la longueur moyenne d'un chemin d'accès à une page. Ce type de connaissance, aide à améliorer la performance des sites web, à augmenter la sécurité globale du système, à faciliter la maintenance du site, et à fournir une aide pour des décisions commerciales (marketing support).

Règle d'association La technique basée sur les règles d'association est une technique destinée à découvrir des associations ou des corrélations intéressantes , parmi un grand ensemble de données (Tianyi, 2001), et élaborer ainsi des prévisions basées sur le principe de co-occurrence. Dans le domaine du Web, des règles d'association sont employées pour indiquer des corrélations entre les pages accédées ensembles pendant les différentes sessions. De telles règles indiquent des rapports possibles entre les pages qui sont souvent consultées ensembles, même

si elles ne sont pas directement connectées (liées). Ces corrélations peuvent aussi indiquer des relations entre groupes d'utilisateurs partageant certains intérêts.

Groupement "Clustering" Il s'agit d'une technique destinée à grouper ensemble les éléments ayant des caractéristiques semblables. Le but général du clustering est de séparer un groupe de données quelconques en groupes appelés Cluster tels que les éléments du même cluster soient similaires entre eux et dissimilaires des autres cluster. Il s'agit d'une méthode intéressante car le clustering fonctionne le plus souvent en mode non supervisé c'est-à-dire que l'analyse ne fournit pas *a priori* les propriétés définissant les groupes.

Classification La classification consiste à classifier les données élémentaires dans une ou plusieurs classes prédéfinies (Fayyad et al., 1996). Dans le domaine du Web la classification est employée pour caractériser des profils d'utilisateurs à partir de leur classe d'appartenance. Par exemple la classification sur fichiers d'accès log de WWW peut mener à l'extraction d'informations comme le fait que 60% des clients qui ont passé une commande en ligne sur /company/produits/produit7, étaient dans le groupe d'âge 27-30. La classification peut être basée sur les arbres de décision, les réseaux de neurones (Benedek et Trousse, 2003), les réseaux Bayésiens (Breese et al., 1998), etc...

Découverte de motifs séquentiels Il s'agit d'une prolongation des règles d'association parce qu'elle est basée sur des modèles de cooccurrence mais elle incorpore également la notion d'ordre. Dans le domaine du Web un tel modèle est constitué d'un ensemble de pages Web accédées les unes après les autres. L'objectif des techniques fondées sur des motifs séquentiels (sequential patterns) (Mannila et al., 1995; Srikant et Rakesh, 1996) est d'extraire toutes les sous-séquences les plus fréquentes. A partir de ces modèles séquentiels les applications web cherchent à prévoir les navigations probables de l'utilisateur et à personnaliser l'environnement comme les fenêtres publicitaires liées à chaque groupe d'utilisateurs.
Les différentes applications et approches existantes dans le domaine du Web Usage Mining présentent des avantages et des inconvénients qui sont souvent très liés au type d'application choisie. Certaines prennent en compte le temps de navigation, d'autre non, certaines s'intéressent au nombre de classes obtenues etc...Du fait de ces particularités liées aux domaines d'application ces méthodes sont en général difficilement adaptables à d'autres domaines.

En se basant sur les avantages et inconvénients des méthodes présentées, nous avons conçu notre propre approche. Elle s'appuie sur une fonction de similarité basée sur un algorithme de recherche séquentielle et un modèle de "prédiction n-gram" afin de trouver les parcours prototypique pour un site web. Même si notre objectif à long terme est la modélisation de l'opérateur, cette analyse ne visait aucun objectif particulier autre que celui de la découverte de ces "patrons de navigation". Cette ouverture permet d'envisager l'application de notre approche à un grand nombre de domaines d'utilisation du Web Mining.

3 Extraction De Patron De Navigation

En se rapportant à beaucoup d'approches existantes et en étudiant les besoins des différentes applications de Web Usage Mining nous avons tiré certaines contraintes qui nous ont

guidé dans notre approche pour la recherche de patron de navigation. Extraire un patron de navigation implique une phase de regroupement des sessions utilisateurs en fonction de leur proximité. Ce regroupement implique la recherche, pour une séquence donnée, des séquences similaires à celle-ci. Une des spécificités de notre travail tient dans la volonté d'assimiler des séquences qui ne sont pas absolument identiques. Il s'agit de ne pas tenir compte de ce que l'on peut considérer comme erreur négligeable dans le parcours de l'internaute, et de différencier les recherches réellement différentes.

3.1 Mesure De Similarité Entre Parcours

La constitution des groupes qui servent de support à la définition d'un parcours prototypique est basée sur une fonction de similarité qui fournit un taux de similarité pour un couple de sessions donné. L'extraction des sessions, puis des parcours se fait, de manière très classique, à partir des fichiers journaux (fichiers log) du serveur d'un site web donné. Pour produire une valeur de ressemblance, cette fonction de similarité prend en compte un certain nombre de facteurs dont les plus importants sont les suivants :

- Ordre des pages dans la séquence : nous considérons qu'il s'agit d'un facteur essentiel révélateur de deux comportements semblables de navigation. Il s'agit de l'ordre d'apparition des pages dans la session. Par exemple : la séquence "P1 P2 P3 P4" sera considérée n'est "pas similaire" à la séquence "P1 P3 P2 P4", même si les mêmes pages ont été visitées dans ces deux séquences.

- Tolérance aux erreurs : capacité de négliger des pages qui peuvent être considérées comme "visitées par erreur". A titre d'exemple, faut-il considérer les deux séquences seq1 : P1 P2 P6 P3 P4 et seq2 : P1 P2 P3 P4 différentes ou pas ? La réponse que nous proposons est une réponse graduelle, où une telle différence sera ou non négligée selon un paramètre de la fonction, et selon certaines autres caractéristiques comme le nombre de pages différentes relativement à la longueur des séquences.

- Comparaison de séquences de longueur différentes : même si les séquences présentent des longueurs différentes, il peut être utile de déterminer que certains parcours ne sont en fait que des sous-parcours d'autres plus grands, mais dont la finalité est identique. Pour ce faire nous avons utilisé un modèle de prédiction n-gramme, qui et après avoir constaté la similarité sur les premiers éléments, vérifiera si les séquences pourraient être identiques. Il s'agit de déterminer s'il est très probable ou pas que la séquence la plus courte, soit en conformité avec la séquence la plus longue. Par exemple : dans seq1 : P1 P2 P3 et seq2 : P1 P2 P3 P4, nous considérerons que la séquence P1 P2 P3 est similaire à la séquence P1 P2 P3 P4 si, d'après les autres sessions, la page P4 est une page très probable après une séquence P1 P2 P3. Ce modèle de prédiction est fondé sur le "modèle n-gramme" (Jelinek, 1998).

Ces facteurs associés présentent la nouveauté et les avantages de notre approche. Par exemple, un des avantages de notre méthode est de conserver l'intégralité des séquences utilisées dans l'analyse. Les prototypes qui représentent les classes sont nécessairement des parcours existants. Les approches statistiques ou les autres approches de Data Mining, ne s'imposent pas en général cette contrainte car cela ne présente pas d'intérêt dans le cadre de ces approches. Dans notre cas ce choix s'explique entre autres par notre objectif à moyen terme : reconstruire le processus cognitif mis en oeuvre par l'utilisateur dans sa recherche. Or dans cette perspective nous faisons l'hypothèse que chaque étape de la recherche est importante car elle vient modifier les connaissances de l'utilisateur et modifier aussi sa requête (en fonction des nou-

velles connaissances acquises). Dans ce contexte chaque étape à un impact potentiel qui sera représenté dans le modèle du processus de recherche.

3.2 Mise En Oeuvre De La Fonction

On peut diviser le traitement réalisé par la fonction de similarité en trois parties : Recherche, vérification, filtrage (FIG. 1).

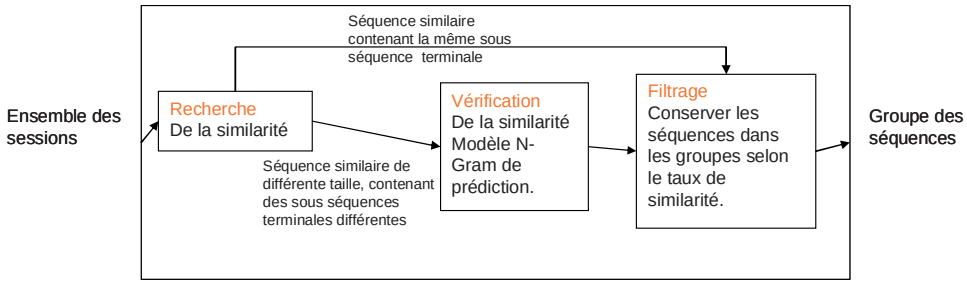


FIG. 1 – *Structure de la fonction de similarité*

La fonction de similarité reçoit en entrée les séquences filtrées provenant des fichiers log du site. La procédure d'extraction des patrons produit comme sortie, des groupes disposant chacun d'un représentant prototypique comprenant une séquence avec laquelle tous les membres ont été jugés similaires.

3.2.1 Recherche

La partie "recherche" de la fonction de similarité consiste à déterminer la similarité entre toutes les séquences reçues à l'entrée. Cette étape est paramétrable avec les facteurs suivants : Couverture, Erreur, TauxErreur, SeqMarge. Chacun de ces paramètres ajoute une règle au calcul de similarité. Dans le cas oú tous ces règles sont vérifiées, on passe à l'étape suivante.

Couverture : Positionne le nombre minimal souhaité d'éléments communs entre les deux séquences. C'est un pourcentage qui par défaut est égal à 75%.

Exemple : Etant donné seq1 = P1 P2 P3 P4, seq2 = P1 P2 P3 et Couverture = 90%. Il faut que 3.60 pages $\simeq$ 4 pages par ordre de seq1 apparaissent dans seq2 pour qu'elles soient similaires. Et dans ce cas on considère que seq2 n'est pas similaire à seq1 (car il existe moins que 4 éléments de seq1 dans seq2 (3 éléments P1 P2 P3)).

Erreur : Ce paramètre nous permet de considérer comme négligeable un ensemble d'actions entreprises par l'utilisateur entre deux actions données. On positionne ici le nombre maximum d'actions que peut inclure cet ensemble que nous qualifions "d'erreur". Ce paramètre indique le nombre maximum de pages visitées par erreur présentes entre deux pages successives de la séquence. Par défaut ce paramètre est égal à 1. Exemple : Si Erreur=1 et que l'on compare seq1 = P1 P2 P3 à seq2=P1 P7 P2 P3 ou seq3= P1 P7 P2 P8 P3 on obtiendra un résultat positif

dans les deux cas. Une séquence telle que seq4= P1 P7 P9 P2 P3 donnera par contre un résultat négatif.

TauxErreur : Corrélé au précédent, ce paramètre permet de relativiser l'importance de l'erreur par rapport aux tailles des séquences évaluées. Ce paramètre est un taux servant à prendre en compte l'importance de l'erreur par rappport à la taille totale de la séquence. C'est un pourcentage, par défaut égal à 15%. Ce paramètre est en lien direct avec le nombre d'élément en commun entre les 2 séquences. (nep =TauxErreur*max/100 , avec <max> la longueur maximale de sous-séquences en commun entre les 2 séquences et <nep> le nombre maximum permis d'erreur). A titre d'exemple, si l'on considère les deux séquences seq1 : P1 P2 P3 et seq2 : P1 P10 P2 P6 P3 alors on rejettera la similarité dans ce cas. En effet, 2 erreurs pour 3 pages communes fournissent un taux d'erreur supérieur à 15%.

SeqMarge : ce paramètre nous permet de tenir compte de la différence de longueur entre les séquences à comparer (c'est-à-dire nombre d'éléments qui apparaissent et qui ne sont ni erreur ni élément de la séquence). Ce paramètre est une valeur entière qui est par défaut égale à 1, et qui sera multiplié par le nombre maximum permis d'erreur. A titre d'exemple, on a SeqMarg=1 et Erreur=1, comparons les deux séquences seq1 : P1 P2 P3 et seq2 : P10 P1 P6 P2 P3 P12 P13. La sous-séquence P1 P2 P3 (3 éléments) est incluse dans seq2 avec une page d'erreur P6. Les 3 autres éléments restants P10, P12 et P13 ne sont ni des pages naviguées par erreur ni des pages de la séquence en commun avec seq1.
Pour savoir si la marge est acceptée on calcul :
nombre_des_pages_maximum_à_la_marge = SeqMarge$\times$nb_maximum_d'erreur_permis = 1×1 = 1 élément. Or il existe 3 éléments (P10, P12, P13) alors on considère que seq2 n'est pas similaire à seq1.

3.2.2 Vérification

L'étape de vérification est utilisée pour vérifier la similarité entre les séquences de tailles différentes. Un modèle de prédiction n-gramme est utilisé afin de vérifier si on pourra compléter les séquences pour qu'elles deviennent de même taille.

3.2.3 Filtrage

Dans cette dernière étape, il s'agit d'éliminer les groupes ayant un nombre d'éléments inférieurs à un nombre fourni par l'utilisateur. Ensuite on filtre ces groupes de manière à ce qu'une séquence donnée ne soit présente que dans un seul groupe, celui où elle possède la plus grande valeur de similarité. Dans le cas oú la valeur de similarité d'une séquence donnée est la même dans plusieurs groupes, on la laisse dans tous les groupes.

4 Résultat

Les résultats de la fonction de similarité sont des groupes identifiées par des prototypes qui représentent les parcours les plus représentés dans le fichier log. Cette liste des parcours prototypiques et de groupes sera considérée comme un mini fichier log où nous pouvons appliquer

n'importe quelle méthode statistique pour obtenir par exemple : la page la plus visitée dans le site etc…
Selon nos tests sur les fichiers log de deux sites différents, fichiers du site du laboratoire LSIS et du site de la formation MIAGE, nous avons obtenu une liste des groupes identifiée par des parcours prototypiques significatifs. Par exemple dans le cas du site du laboratoire LSIS le parcours prototypique le plus utilisé représentant le plus grand groupe des séquences est : P3 → P15→ P22→ P1 qui correspond à (/ → ' Equipes '→membres '→ Fiche) ce qui est cohérent dans le contexte d'un site de laboratoire. En analysant les résultats on a constaté que le nombre de groupe varie entre un fichier et un autre en conservant toutefois certains prototypes en commun. Ce nombre de groupes tend à devenir constant quand le nombre de sessions augmente. (FIG. 2).

Fichiers Log	319 sessions	380 sessions	3613 sessions
Nb de groupes	14 groupes	20 groupes	94 groupes
Nb de groupes en commun	~= 7 groupes		
Le Parcours le plus intéressants	=	=	=

FIG. 2 – *Evaluation des résultats selon le nombre de sessions*

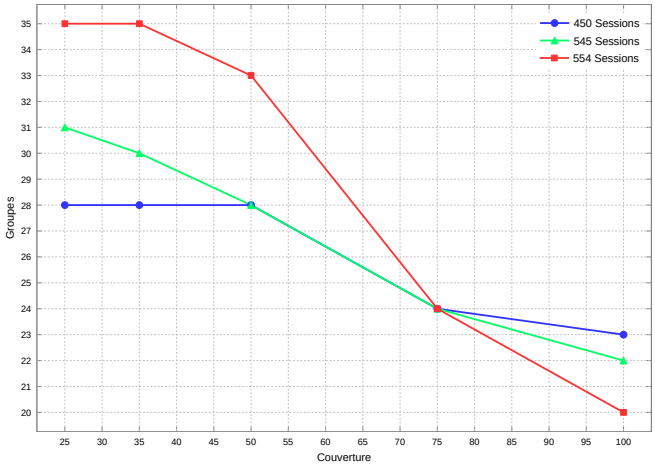


FIG. 3 – *Variation du paramètre couverture*

Actuellement nous examinons la possibilité de développer un système de recommandation où les prototypes constitueront notre base de données de comportements. Nous pouvons utiliser n'importe quelle méthode de prédiction basée sur les probabilités et/ou un modèle de prédiction (exemple n-gramme (Jelinek, 1998)) en exploitant cette base de données, afin de prédire les choix futurs de l'utilisateur et de l'orienter dans sa recherche.

Nous avons testé l'efficacité et la stabilité de nos paramètres sur différents fichiers log. En faisant varier les valeurs des différents paramètres, nous avons constaté que le nombre de groupes résultant varie d'une manière décroissante avec l'augmentation de la valeur de la couverture (FIG. 3). Ceci vérifie la bonne fonctionnalité de ce paramètre, puisque l'augmentation de sa valeur augmente le caractère discriminant de la comparaison entre deux séquences.
De même les autres paramètres ont été vérifiés et testés de la même façon.

5 Conclusion

S'il existe beaucoup d'outils d'analyse statistique des visites de site web (page la plus visitée, temps moyen de visite, etc. . .), il existe moins d'outils qui s'intéresse à la cinématique de navigation, pourtant riche d'enseignement sur l'internaute. Nous avons proposé une procédure et une fonction de similarité afin d'extraire des patrons de navigation en explicitant les critères de regroupement et de sélections des sessions. Une des particularités de notre approche est d'offrir une certaine tolérance aux erreurs de navigation. En effet, nous faisons l'hypothèse qu'un internaute peut chercher son chemin sur le site et par conséquent emprunter des variantes qui ne remettent pas en cause son parcours global. Les résultats obtenus sont satisfaisants. Réalisés sur plusieurs sites à partir de différents fichiers journaux, ils révèlent des pratiques de navigation cohérentes par rapport au contenu du site et aux statistiques statiques. La question que nous étudions maintenant peut être formulée en ces termes : les patrons que nous extrayons peuvent-ils être considérés comme correspondant à une certaine tâche de recherche d'information ? Sont-ils différenciées seulement par leurs buts ou témoignent-ils aussi de la structure d'une tâche de RI ? Les perspectives actuelles portent sur la modification de notre outils de recommandation selon les réponses que nous produiront pour ces questions.

Références

Aggarwal, C. C., W. Joel L., et Y. Philip S. (1999). Caching on the world wide web. *Knowledge and Data Engineering 11*(1), 95–107.

Baeza-Yates, R. (2004). Web mining in search engines. In *Proceedings of the 27th Australasian conference on Computer science - Volume 26*, Dunedin, New Zealand, pp. 3–4.

Baoyao, Z., C. H. Siu, et C. M. F. Alvis (2005). Mining longest repeating subsequences to predict world wide web surfing. In *Proceedings of the Workshop on Personalization on the Semantic Web (PerSWeb 05)*.

Baoyao, Z., C. H. Siu, et C. Kuiyu (2004). An intelligent recommender system using sequential web access patterns. In *Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems*, Sch. of Comput. Eng., Nanyang Technol. Univ., Singapore, pp. 1–3.

Benedek, A. et B. Trousse (2003). Adaptation of self-organizing maps for case indexing. In *27th Annual Conference of the Gesellschaft fur Klassifikation*, Cottbus, Germany. Improved version from one in Proceeding of the 4th International Workshop on Symbolic and Numeric Algorithms for Scientific Computing, Vol. XL, Special Issue on Computer Science :35-52, Timisoara, Romania, October 2002.

Breese, J. S., H. David, et K. Carl (1998). Empirical analysis of predictive algorithms for collaborative filtering. Technical report, Microsoft Research, One Microsoft Way, Redmond, WA 98052.

Cohen, E. et K. Haim (2000). Prefetching the means for document transfer : A new approach for reducing web latency. In *INFOCOM (2)*, pp. 854–863.

Cooley, R., M. Bamshad, et S. Jaideep (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems 1*(1), 5–32.

Fayyad, U. M., P.-S. Gregory, et S. Padhraic (1996). From data mining to knowledge discovery in databases. *Ai Magazine 17*, 37–54.

Jelinek, F. (1998). *Statistical Methods for Speech Recognition*. MIT Press.

Mannila, H., T. H., et V. A. I. (1995). Discovering Frequent Episodes in Sequences. In U. M. Fayyad et R. Uthurusamy (Eds.), *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, Montreal, Canada. AAAI Press.

Perkowitz, M. et E. O. (2000). Towards adaptive web sites : Conceptual framework and case study.

Pitkow, J. E. et P. Peter (1999). Mining longest repeating subsequences to predict world wide web surfing. In *USENIX Symposium on Internet Technologies and Systems*.

Sarukkai et R. R. (2000). Link prediction and path analysis using markov chains. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking*, Amsterdam, The Netherlands, The Netherlands, pp. 377–386. North-Holland Publishing Co.

Shim, J., S. Peter, et V. Radek (1999). Proxy cache algorithms : Design, implementation, and performance. *Knowledge and Data Engineering 11*(4), 549–562.

Srikant, R. et A. Rakesh (1996). Mining sequential patterns : Generalizations and performance improvements. In G. G. Peter M. G. Apers, Mokrane Bouzeghoub (Ed.), *Proc. 5th Int. Conf. Extending Database Technology, EDBT*, Volume 1057, pp. 3–17. Springer-Verlag.

Srivastava, J., C. Robert, D. Mukund, et T. Pang-Ning (2000). Web usage mining : Discovery and applications of usage patterns from web data. *SIGKDD Explorations 1*(2), 12–23.

Tianyi, L. (2001). *Web-Document Prediction And Presending Using Association Rule Sequential Classifiers*. A thesis submitted in partial fulfilment of the requirements for the degree of master of science, Simon Fraser University.

Summary

This paper describes an approach for automatically finding prototypic navigation behaviour of web users. User access log files are examined in order to group users that have a similar access pattern with a sequence pattern and to extract frequent sequences. This result gives us an efficient way to better understand the way users are acting, and lead us to improve the structure of website for navigation convenience. Another interesting application where this function can be used is in recommender systems, we can use it to suggest for users links that could interest them. In this paper we describe the overall design of a similarity function that presents all these features, and we will introduce you some of our results.

Visualisation personnalisée d'un corpus de conférences scientifiques

Romain Vuillemot*

*LIRIS, INSA Lyon, F-69621, Lyon, France,
romain.vuillemot@insa-lyon.fr,
http://liris.cnrs.fr/romain.vuillemot/

Résumé. Dans cet article nous proposons une interface de visualisation personnalisée d'un site Web regroupant des conférences scientifiques de manière structurée sous forme d'un corpus. Nous proposons une interface de visualisation multi points de vue combinant entre autres lieux, dates et mots clé de manière synchronisée. L'utilisateur peut choisir, personnaliser et agencer ces vues dans le plan du navigateur Web. Nous proposons également des méthodes de réduction des données par la création de vues contextuelles et le filtrage dynamique local au client. Les préférences utilisateurs sont captées à la fois de manière implicite et explicites, puis stockées dans un profil utilisateur. Un prototype a été réalisé et les premiers résultats sont présentés.

1 Introduction

Une des nombreuses tâches de tout chercheur consiste à produire et communiquer des connaissances, notamment au travers de conférences organisées un peu partout dans le monde. Cependant, face à leur multiplication et à l'effort de pluridisciplinarité actuel, il est extrêmement difficile de pouvoir effectuer une veille efficace de tous ces événements.

L'information relative aux conférences est disponible sur le Web, mais à notre connaissance il n'existe pas d'outil efficace de centralisation, formalisation et représentation interactive de ces informations. Et encore moins un outil intégrant les préférences de l'utilisateur pour offrir un accès adapté à ses caractéristiques, bien que des problèmes similaires ont déjà fait l'objet d'études (Rossi et al. (2001)). Le jeu de données de conférences existe donc, il est bien connu de tout chercheur qui s'en fait une représentation mentale implicite, mais à l'heure actuelle il n'est pas instrumentalisé. Ainsi les représentations actuelles des conférences sont sous la forme de la figure 1 qui, 1) affiche des données structurées sans sémantique et 2) utilise photos, couleurs et symboles qui reprennent les codes de la conférence et renforcent le modèle mental de l'utilisateur. Ce sont des informations qui ne permettent pas le traitement automatique par la machine. Il est donc nécessaire de formaliser le jeu de données sous-jacent, et ensuite proposer un accès efficace à ces données selon les caractéristiques de chaque utilisateur afin de proposer une interface interactive remodelée selon ses préférences (Anderson et Horvitz (2002)).

FIG. 1 – *Le site de l'atelier sur la* Modélisation Utilisateur et la Personnalisation d'Interfaces Web (MUPIW) *de la conférence* Extraction et Gestion des Connaissances (EGC) 2008 *est typique, de part son design attrayant mais également de part son manque de sémantique.*

1.1 Le jeu de données corpus de conférences

Par analyse d'échantillons représentatifs de nombreux sites existants, nous avons constaté que les conférences sont structurées et visualisables de trois façons : par leur *description structurée* (nom de conférence, numéro d'édition, ..), *composante spatio-temporelle* (lieu et date) et *autres composantes* (mots-clés, réseaux sociaux, annotations de l'utilisateur, ..). En voici les détails :

Description structurée : la description d'une conférence est une suite de données explicites toujours présentes sur les sites Web de conférences. Elles sont composées et structurées selon un schéma récurrent : d'abord le titre, ensuite le numéro de l'édition, la date, le lieu, etc... Ainsi nous proposons de structurer les conférences selon ce schéma là, et l'ordre sera repris pour l'affichage de ces descriptions sous forme de texte.

Composante spatio-temporelle : chaque conférence possède une composante spatio-temporelle, à savoir un lieu associé à une date (ou réciproquement). Concernant la composante *temporelle*, il peut s'agir du début ou de la fin de la conférence, ou des dates limites (qui peuvent être un moment précis ou une durée). Enfin la composante *spatiale* indique le lieu de déroulement de la conférence.

Autres composantes : ce sont des données non-présentes sur tous les sites (comme les mots-clés par exemple) ou qui apparaissent par l'instrumentalisation technologique de l'accès aux conférences (apparition de données sociales par exemple via un site unique qui regroupe tous les accès).

A partir de ce jeu de données, nous allons, dans une deuxième partie, voir quelles en sont les représentations possibles. Dans une troisième partie nous verrons comme les choisir et les combiner de manière personnalisée. Une quatrième partie évoquera l'acquisition et le stockage des préférences dans un profil utilisateur. Enfin dans une cinquième partie les détails du prototype développé seront évoqués, et dans une sixième partie les conclusions et perspectives de notre démarche.

2 Vues multiples interactives sur le corpus de conférences

Une bonne représentation visuelle sera celle qui correspondra le mieux à celle que s'en fait déjà un utilisateur dans son esprit. Cependant, étant donné le nombre d'attributs des données, il est indispensable de les représenter et de les composer séparément dans l'espace 2D du navigateur, puis d'interagir avec elles de manière coordonnée.

2.1 Multiplicité des vues

Le premier type de données à visualiser est la *description structurée* des conférences. Comme annoncé en introduction, nous reprenons la structure déjà existante sur les sites des conférences ou les appels à participation par courrier électronique sous forme de texte. Dans le cas d'affichages de plusieurs conférences, nous avons extrait les informations de nom, d'acronyme et date, et disposées sous forme de liste linéaire (figure 3). Cette liste peut être réordonnée par utilisateur, selon la date ou le nom.

Le second type de donnée est la *composante spatio-temporelle* des conférences. Pour les informations relatives au temps, nous proposons Timeline[1] qui consiste en plusieurs bandes chronologiques à différentes échelles (jour, mois, an) permettant de voir ce qui se passe avant et après. D'autres vues comme le calendrier sont envisageables. Pour les informations géographiques, nous avons utilisé Google Map[2] sur laquelle sont affichées les conférences par icône ou symbole. Des actions de zoom, déplacement et demande de détails sur la carte sont possibles.

Enfin le troisième type de données et le plus délicat : il s'agit de la visualisation des *autres composantes*. En effet, les données énumérées dans les deux paragraphes précédents ont un support explicite de représentation. Par exemple les villes sur une carte géographique qui représente pays et continents, et les dates sur un calendrier. Dans le cas de mots-clés ou de réseaux sociaux, il s'agit de relations indépendamment d'un lieu ou d'une période. Ainsi il faut regarder vers le domaine de la visualisation d'information (infovis) pour y trouver des contributions intéressantes. Cependant il n'existe pas de méthode automatique de représentation de données relationnelles et ce domaine est très actif. Il s'agit donc d'implémenter des métaphores ou d'installer des librairies pour bénéficier des dernières contributions[3]. Nous verrons dans la section 3 que l'utilisateur aura le choix parmi ces représentations visuelles. Pour les mots-clés nous pouvons par exemple proposer les *Tag Cloud* [4] qui permettent aux mots d'occuper un espace et d'intégrer l'attribut fréquence par leur taille ou couleur. Pour

[1] http://simile.mit.edu/timeline/
[2] http://www.google.com/apis/maps/
[3] http://www.cs.umd.edu/hcil/InfovisRepository/
[4] http://en.wikipedia.org/wiki/Tag_cloud

représenter les réseaux sociaux nous pouvons utiliser des librairies de représentation de graphes telles que LGL (Adai et al. (2004)), basé sur un formalisme type FOAF[5]. D'autres vues sont envisagées comme des cartographies de réseaux de co-publication ou de co-citation.

2.2 Interactivité des vues

Chacune des vues représente des attributs d'informations relatives à une conférence ou un ensemble de conférences. Il est désormais nécessaire de trouver une méthode efficace d'interaction avec les données permettant d'augmenter les capacités cognitives de l'utilisateur (North et Shneiderman (2000)). Nous proposons deux approches : des *interactions locales* au client et valables pour une page, et des *interactions globales* qui seront une vue restreinte de la totalité des données et la même pour toutes les pages.

Interactions locales : elles permettent d'avoir plus de détails au travers de chacune des visualisations présentées dans la sous-section précédente. Par exemple le zoom sur la carte permet de se focaliser sur une conférence, de cliquer sur son nom et de passer à sa page de détails. Nous proposons également le filtrage dynamique qui permet de n'afficher dans la page que les informations possédant un attribut contenant le mot-clé, saisi par l'utilisateur une fois la page chargée localement. Par exemple l'entrée du mot `Gestion` n'affichera que les conférences qui ont un nom, lieu ou date contenant `Gestion`. Les interactions locales permettent de faciliter la recherche dans de nombreux champs sans avoir à effectuer de transferts de données avec le serveur. Cela permet une plus grande réactivité aux requêtes ainsi qu'une tolérance plus élevée aux erreurs de l'utilisateur. Une autre possibilité de filtrer existe si l'utilisateur zoome sur un pays, les données affichées sur la page seront réduites à celles visibles sur la carte. Ainsi les dates affichées dans la Timeline seront celles concernant les conférences ayant lieu et ayant eues lieu dans ce pays. Nous retrouvons une forme de filtrage dynamique mais cette fois plus intuitive, notamment dans le cas où l'utilisateur ne sait pas utiliser les codes du système. Par exemple si on ne connaît pas l'orthographe d'un pays on peut zoomer dessus et les informations affichées seront liées à ce pays même si l'utilisateur ne l'a pas entré directement.

Interactions globales : elles permettent la mise en place d'un contexte de navigation. Cela n'est pas nécessaire dans une phase exploratoire où la surcharge d'information permet la recherche visuelle. Mais si l'on connaît déjà bien les conférences qui nous intéressent, il est nécessaire de restreindre le jeu de données à une vue sur la base. En d'autres termes il s'agit de définir explicitement un contexte d'utilisation, qui sera activable ou désactivable (plus de détails dans la section 4) au gré de l'utilisateur.

3 Choix et agencement personnalisé des vues

L'utilisateur a désormais en main plusieurs vues intelligibles sur les données et certaines lui paraissent plus pertinentes que d'autres. Cependant, son espace de travail étant toujours limité par la fenêtre du navigateur, la prochaine étape sera donc *le choix des points de vue*, et ensuite *leur agencement dans le plan*.

[5] Friend Of A Friend : http://xmlns.com/foaf/spec/

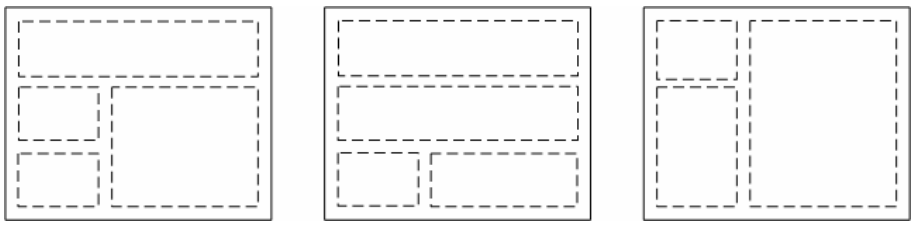


FIG. 2 – *Des patrons de mise en page sont déjà défini et l'utilisateur n'a plus qu'à choisir quelle vue, et la glisser déplacer au bon endroit (dans le cadre en point tillés).*

Choix des vues : il permet à chaque page une mise en perspective spécifique des données. Le choix des vues permet également de naviguer dans une vue plutôt qu'une autre, comme la navigation dans les dates par année, mois et jours. Ainsi une vue devra sans doute prédominer dans l'espace de navigation par la taille et la position de sa fenêtre. Les vues de *description* et de *composantes spatio-temporelles* sont maîtrisées par tout utilisateur, mais d'autres le sont moins comme la vue en nuage de mots-clés. Ces vues pourront donc être marginalisées dans l'espace le temps de leur familiarisation.

Agencement selon des patrons de conception : trop de liberté dans la personnalisation de la disposition des vues dans le plan peut nuire à l'intelligibilité de l'interface. En effet, le positionnement des fenêtres doit suivre un ordre de positionnement qui ne doit pas laisser d'espaces vides, comme cela se produirait avec liberté de positionnement totale (ou même avec un support de grille). De surcroît, tout utilisateur ne possédant pas les compétences d'un ergonome ou d'un graphiste, nous lui proposons de choisir parmi des cadres préconçus dans lesquels l'utilisateur déplace la vue au bon endroit du cadre (figure 2). Le patron reste le même pour chaque page pendant toute la navigation sur le site. Mais l'échelle, le niveau de granularité ou d'agrégation de chaque vue varie selon la précision des informations. En prenant pour exemple les données géographiques, si l'on affiche toutes les conférences d'un domaine, les données sont d'ordre mondial donc une carte du monde est affichée. Cependant pour les données relatives à l'édition d'une conférence on peut afficher le plan détaillé de la ville.

Ceci dit, une approche mixte est parfois nécessaire : on veut parfois afficher à la fois une vue globale et une vue détaillée. Nous proposons deux solutions :

1. Dans le cas des périodes, si nous affichons à la fois les horaires des sessions d'une conférence sur la Timeline, mais aussi les appels à participation par effet Fisheye (Furnas (1986)) afin de garder une information contextuelle.
2. Dans le cas du lieu d'une conférence, nous affichons le plan détaillé d'une ville mais il y a alors perte de l'information relative au nom du pays. Pour cela nous ajoutons une mini-carte en bas à droite de la carte principale, à un niveau de zoom plus élevé qui est à l'échelle du pays cette fois-ci.

4 Personnalisation

Face d'un côté à la quantité importante d'informations et d'un autre à ses nombreuses possibilités de représentation et d'agencement, un choix est nécessaire. Nous avons donc mis en place une approche de personnalisation mixte du site Web, à la fois de manière *explicite* par l'utilisateur (qui choisi par exemple ses conférences préférées), mais aussi *implicite* par le système (par étude des comportements récurrents). Au moyen de ces deux approches, nous allons dans un premier temps collecter les informations sur l'utilisateur, ensuite étudier la construction de contextes de navigation et enfin nous verrons la structure du profil utilisateur.

4.1 Collecte des informations

Avant de commencer la collecte, il est nécessaire d'identifier l'utilisateur (Brusilovsky et al. (2007)). Tout utilisateur, connu ou inconnu, sera identifié par une variable de session stockée dans un cookie. Cette méthode est non-intrusive et supportée par tous les navigateurs. Si l'utilisateur a un compte, nous lui proposons de s'identifier et il pourra ainsi retrouver ses préférences (contextes de navigations sauvegardés, consultation de son historique, ..) et modifier ses informations personnelles. La collecte se déroulera de deux façons, selon si les données générées le sont de manière explicite ou implicite :

Données explicites. Elles sont collectées via l'action utilisateur qui clique sur un drapeau (bouton) permettant d'indiquer ses conférences favorites, celles où il va publier ou encore les personnes avec qui il travaille. Par ce moyen, l'utilisateur pourra créer une communauté et le système lui suggérera des conférences potentiellement d'intérêt, car préférées d'autres personnes de sa communauté. Un tableau de bord sera proposé à l'utilisateur qui pourra retrouver toutes les traces de ses actions, ce qui permettra par exemple d'afficher les prochaines dates limites de conférences ou recevoir toute extension de date parmi ses conférences où il compte soumettre. Enfin du point de vue de l'interface, le choix des vues et de l'agencement sont considérées comme explicites et automatiquement stockées dans son profil.

Données implicites. Elles sont collectées de manière non-intrusive par le système. Par exemple si l'utilisateur visite certaines conférences alors celles-ci apparaîtront dans une liste « *conférences les plus visitées* ». Cet affichage de fréquentation peut être réduit à l'utilisateur, mais étendu à une communauté ou à tout le système. Et cela devient un système de recommandation. D'autres méthodes d'analyse de comportement, telles que la recherche de traces similaire est envisageable, mais demande une analyse *hors-ligne* (Eirinaki et Vazirgiannis (2003)), qui n'est pas effectuée en temps réel.

Ces deux types de données seront par la suite stockés dans le profil de l'utilisateur. A noter que la confidentialité de l'utilisateur est garantie par la fiabilité technique du site Web, mais aussi par le choix qu'il possède dans la diffusion ou non d'informations sur son profil.

4.2 Contexte de navigation

L'utilisateur peut restreindre sa navigation dans le jeu de données à un sous-ensemble de conférences, mots-clés ou lieux qui l'intéressent à un moment donné. Ainsi la construction de contextes de navigation permet de limiter le jeu de données à l'état mental dans lequel est l'utilisateur, tel que: « *Je ne veux m'intéresser qu'aux conférences A+ répondant au mot-clé* `Connaissances` *à Hawaï, de juin à août 2008!* ». Une interface visuelle de construction de requête est proposée sous forme de formulaire, et il est possible d'y associer une description, permettant la sauvegarde pour une réutilisation future ou un partage au sein d'une communauté. L'utilisateur peut ensuite naviguer dans ce contexte de données, le supprimer ou l'augmenter dynamiquement. Ce contexte est activable et désactivable très rapidement et permet ainsi d'effectuer des recherches propres à des thématiques ou objectifs de publication différents. Le nom du contexte activé est affiché en évidence sur la page, ce qui permet à l'utilisateur de ne pas oublier que sa navigation est restreinte à un sous ensemble du jeu de données.

4.3 Profil de l'utilisateur

Le profil utilisateur est constitué de la collection des données précédemment énoncées (sous-section 4.1), ainsi que des contextes de navigation (sous-section 4.2). Il est stocké de manière centralisée, ce qui permet d'y avoir accès quel que soit le lieu où l'utilisateur se trouve. Le profil essaye au mieux de reprendre les intérêts de l'utilisateur, et ces derniers peuvent être distingués en deux catégories: intérêts à court terme, et intérêts à long terme (tableau 1).

Intérêts à court terme	Intérêts à long terme
Edition d'une conférence	Conférence
Contexte de navigation	Domaine scientifique
Destination précise	Proximité géographique

TAB. 1 – *Mise en correspondance des intérêts à court et à long terme.*

La construction du profil est réalisée en deux étapes. La première concerne la partie statique du profil, qui ne change pas au fil du temps, telles que les données personnelles sur l'utilisateur (nom, prénom, âge..). Les autres données (vues précédemment), elles, sont dynamiques car évoluent au fil du temps.

Emergence de stéréotypes. Plusieurs profils types peuvent émerger, par analyse statistique des usages. Cela permettra de constituer des stéréotypes caractéristiques des utilisateurs en termes de choix de patron et de choix des vues. Ainsi tout nouvel utilisateur pourra choisir un profil type et ensuite le décliner selon ses préférences, sans à passer par l'étape de la feuille blanche ou du démarrage à froids.

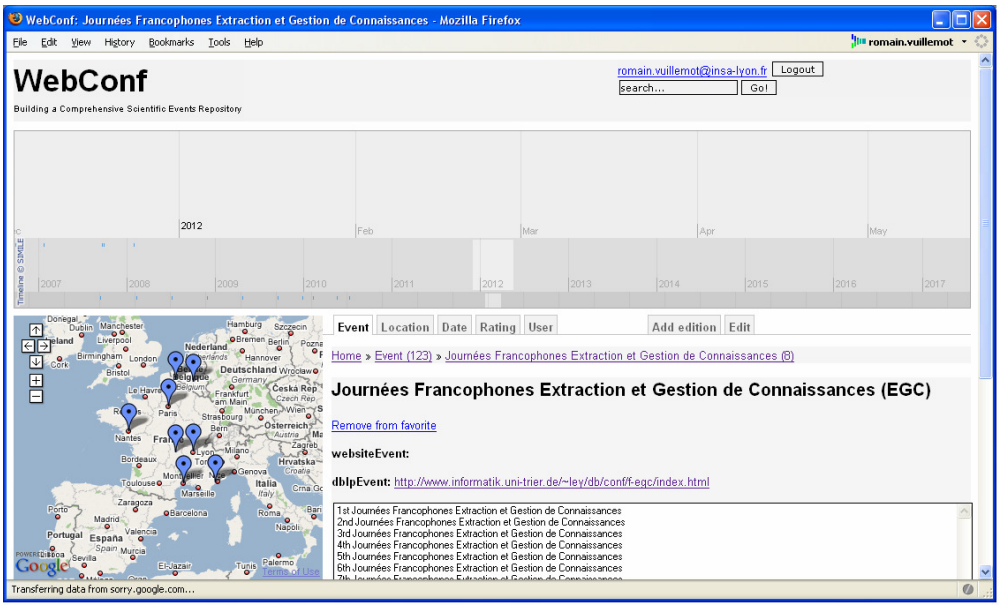


FIG. 3 – *Exemple de navigation dans toutes les conférences EGC organisées. En un coup d'œil on peut déjà réaliser que la conférence est francophone et a toujours lieu en début d'année.*

5 Prototype

Nous avons conçu un site Web opérationnel dont l'interface est présentée sur la figure 3, qui montre par exemple la page regroupant toutes les éditions de EGC (huit au total) depuis la création de la conférence.

Le contexte de navigation est celui des *Journées Francophones d'Extraction et Gestion des Connaissances*. Le patron de la page et le choix des vue offrent tout d'abord une Timeline en largeur de la page. En dessous de la quelle se situe une carte géographique des conférences, et enfin en bas à droite une liste ordonnée des éditions de la conférence.

L'utilisation d'un tel site de manière massive et exhaustive permettra de soulever de nouvelles problématiques, à la fois techniques mais aussi conceptuelles concernant la réduction de masses de données structurées, que nous n'avons pas abordées pleinement dans le cadre de cet article. Les tests seront à effectuer à la fois avec des experts du domaine d'étude (chercheurs expérimentés), mais également avec des personnes extérieurs (non-chercheurs) afin de conserver une utilisabilité ouverte aux non-experts.

6 Conclusion et perspectives

Dans cet article nous avons proposé une interface Web permettant d'accéder de manière personnalisée à un corpus de conférences scientifiques. Nous avons proposé une interface

multi-points de vue, réorganisable par l'utilisateur, ainsi que la prise en compte de préférences utilisateurs et la création de contextes de navigation, stockés dans un profil.

Nous souhaitons pouvoir converger vers une interface visuelle consensuelle (efficace pour un très grand nombre de personnes) d'analyse des données du corpus de conférences, par l'analyse de comportements et de profils récurrents. Cette interface pourra cependant toujours être customisable (c'est à dire subir des modifications mineurs) au gré des préférences utilisateur.

Enfin, nous souhaitons également innover dans l'ajout de représentations visuelles issues du domaine de la visualisation. Ces représentations sont souvent difficiles à maîtriser (sous forme de graphe, réseau, 3D, ..) mais combinées à d'autres plus classiques elles deviennent apprivoisables et permettent d'offrir de nouvelles approches et faire émerger des connaissances nouvelles par analyse visuelle.

Au delà de l'innovation à outrance nous souhaitons aussi mettre en place un accès en mode texte ou à fort contraste pour les personnes à déficientes visuelle.

Remerciements

Merci aux doctorants du LIRIS, pour leur utilisation régulière de notre site de conférences et pour leurs remarques constructives.

Références

Adai, A. T., S. V. Date, S. Wieland, et E. M. Marcotte (2004). Lgl : creating a map of protein function with an algorithm for visualizing very large biological networks. *J Mol Biol 340*(1), 179–190.

Anderson, C. R. et E. Horvitz (2002). Web montage : a dynamic personalized start page. In WWW ’02 : Proceedings of the 11th international conference on World Wide Web, New York, NY, USA, pp. 704–712. ACM.

Brusilovsky, P., A. Kobsa, et W. E. Nejdl (2007). *The Adaptive Web*.

Eirinaki, M. et M. Vazirgiannis (2003). Web mining for web personalization. ACM Trans. Inter. Tech. 3(1), 1–27.

Furnas, G. W. (1986). Generalized fisheye views. In CHI ’86 : Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, pp. 16–23. ACM Press.

North, C. et B. Shneiderman (2000). Snap-together visualization : a user interface for coordinating visualizations via relational schemata. In AVI ’00 : Proceedings of the working conference on Advanced visual interfaces, New York, NY, USA, pp. 128–135. ACM Press.

Rossi, G., D. Schwabe, et R. Guimaraes (2001). Designing personalized web applications. In WWW ’01 : Proceedings of the 10th international conference on World Wide Web, New York, NY, USA, pp. 275–284. ACM.

Visualisation personnalisée d'un corpus de conférences scientifiques

Summary

In this article we introduce a personalized visualization interface of a website, gathering scientific conferences in a structured fashion. We give the outlines of a coordinated multiple viewpoint interface combining –among other- locations, dates and keywords. The end-user has the freedom to personalize its web page layout by selecting and organizing views. We also outline data reduction methods by means of context creation and dynamic local filtering. Users preferences are recorded in both implicit and explicit manner, and all stored in a unique user profile. A prototype has been built and preliminary results are given.

Conception d'un système multi-agent du Web Usage Mining pour la personnalisation du web

Fadoua Ouamani*, Hajer Baazaoui Zghal*,
Zeina Jrad***, Marie-Aude Aufaure**,***, Henda Ben Ghézala*

* Laboratoire Riadi GDL-ENSI-Campus universitaire de la Mannouba
2010-Mannouba, Tunisie
wamanifadoua@yahoo.fr
hajer.baazaouizghal@riadi.rnu.tn
** Supélec - Plateau du Moulon - Service Informatique
91 192 Gif-sur-Yvette Cedex, France
marie-aude.aufaure@supelec.fr
http://www.une-autre-page.html
*** INRIA Paris-Rocquencourt, Domaine de Voluceau
78 153 Le Chesnay Cedex, France
{zeina.jrad, marie-aude.aufaure}@inria.fr

Résumé. Le web usage mining est un outil pour la personnalisation du web qui capture et modélise les comportements et les profils des utilisateurs au cours leurs interactions avec les sites web. Ces modèles peuvent être utilisés par le système de personnalisation pour mieux comprendre et anticiper les comportements des visiteurs des sites web et fournir des recommandations dynamiquement aux visiteurs. L'aspect dynamique du fait qu'il est difficile de cerner tous les aspects du comportement d'un utilisateur du web. Cet article décrit la conception d'un système multi-agent, PWUM, dédié au web usage mining pour la personnalisation du web, PWUM. Le système proposé est composé d'un ensemble d'agents autonomes qui coopèrent afin d'atteindre leurs objectifs principaux. Ce système montre que l'utilisation d'une multitude de techniques de web usage mining en faisant appel aux systèmes multi-agents permet de mieux comprendre et contourner les comportements hétérogènes et instables des utilisateurs du web et leur fournir ainsi une assistance personnalisée et guidée dans le temps requis.

1 Introduction

Avec l'explosion des données disponibles en ligne, le web est rapidement devenu la source principale d'informations pour la majorité des utilisateurs dans plusieurs domaines, ce qui a augmenté le nombre d'utilisateurs inexpérimentés, d'où la nécessité d'analyser et de comprendre les comportements et les intérêts des utilisateurs du web, dans le but de personnaliser les réponses à leurs requêtes. L'objectif de cet article est de répondre à ce besoin en combinant plusieurs techniques de WUM (Web Usage Mining) et en faisant appel à la technologie multi-agent. Nous visons l'amélioration des sorties du processus du WUM en tirant profits du paradigme multi-agent afin de fournir une assistance personnalisée et guidée aux visiteurs des sites web. La technologie agent a été utilisée dans une variété d'applications web comme la recherche d'informations, les web services, le e-learning, le commerce électronique, le web sémantique, le web mining, la personnalisation et l'adaptation du web Gar-

latti & Prié (2003). Ainsi, le paradigme agent est particulièrement conforme aux environnements web distribués, adaptés et personnalisés. Cet article décrit la conception de notre système multi-agent du web usage mining pour la personnalisation du web, PWUM. Le système proposé est composé d'un ensemble d'agents autonomes qui coopèrent afin d'atteindre leurs objectifs principaux. Les agents sont regroupés en modules avec des tâches bien définies et sont scindés en deux groupes de travail : le travail en ligne et le travail hors ligne. Le module de fouille de données est le module le plus important du système, il fournit les patrons au module de personnalisation. Notre système a été implémenté en utilisant la plateforme JADE.

Nous introduisons dans la section 2, les outils et les technologies relatifs aux domaines du WUM et des SMA (Systèmes Multi-agents). Puis, nous présentons notre système et nous décrivons le mécanisme d'interaction entre les différents agents. Ce mécanisme est nécessaire pour assurer le bon déroulement du processus de WUM et les tâches de la personnalisation du web. La section 4 décrit la solution de personnalisation adoptée. Enfin, nous discutons des principaux apports de ce travail, ainsi que des perspectives envisagées.

2 Le web usage mining pour la personnalisation du web

D'après Madria (2007), le WUM est le processus de découverte des patrons et modèles d'usages intéressants, à partir des données d'usage. Ces données sont dérivées des interactions des utilisateurs lors de leurs navigations sur le web Tanasa (2005). Elles décrivent les comportements de navigation des utilisateurs du web et apportent des informations implicites sur les intérêts et les besoins des utilisateurs.

Les sources fondamentales des données d'usage dans le WUM sont les fichiers logs mais on peut collecter des informations d'usage additionnelles Tanase (2005), Pierrakos et al(2003) à partir des cookies Mobasher (2004) et des enregistrements des utilisateurs sur les sites web. Ces sources multiples de données d'usages sont utiles pour la personnalisation du fait qu'elles décrivent les comportements de navigation des utilisateurs, leurs préférences et leurs intérêts. Ainsi, nous pouvons les utiliser dans la modélisation des profils utilisateurs en appliquant des techniques de fouille de données. Les modèles obtenus suite à cette modélisation sont les connaissances opérationnelles pour la personnalisation du web. Dans la sous section suivante, nous décrivons ce processus.

2.1 Le processus du Web Usage Mining

Le web usage mining fournit une approche pour la collecte et le prétraitement des données d'usage et la construction des modèles. Pour la personnalisation du web, le WUM comporte les étapes suivantes :

- la collecte des données : les données d'usage sont collectées à partir d'une seule ou de différentes sources ; leur contenu et leur structure sont ensuite identifiés. Plusieurs techniques et heuristiques peuvent être utilisées dans cette étape Mobasher (2004), Cooley (2000).
- Le prétraitement des données : une fois les données d'usage collectées, elles sont nettoyées, intégrées et traitées dans le but de les préparer comme entrées adéquates et appropriées aux méthodes de fouille de données. Le prétraitement inclut les étapes de filtrage des données, l'identification des utilisateurs et l'identification des sessions utili-

sateurs. Plusieurs heuristiques peuvent être appliquées pour effectuer ces tâches telles que l'heuristique de Cooley (2000).

- La découverte des modèles ou patrons : lors de cette étape, nous appliquons des méthodes de fouille de données sur les données prétraitées (clustering, règles de classification, motifs séquentiels et règles d'association). Ces patrons extraits sont des modèles de comportements et d'intérêts communs aux groupes d'utilisateurs.
- Le post-traitement des connaissances : les connaissances découvertes sont finalement évaluées et incorporées dans le module de personnalisation du web pour essayer d'améliorer les fonctions de personnalisation.

Ainsi, le web usage mining est un outil pour automatiser la personnalisation du web. Des travaux de recherche se sont intéressés à tirer profit à la fois du web usage mining et des systèmes multi-agents pour bâtir des systèmes de personnalisation du web.

2.2 Les systèmes multi-agent pour la personnalisation du web et pour le WUM

Plusieurs travaux et recherches se sont intéressés aux systèmes multi-agent pour la personnalisation web en utilisant comme outil le WUM. WUMA-MAS Girardi et al (2005) est une architecture multi-agent pour la construction d'un système web de recommandation qui traite différents problèmes : la modélisation des utilisateurs, la recherche d'informations, le filtrage et les systèmes de recommandations. SETA Ardissono et Torasso (2000) est également une plateforme multi-agent qui peut être utilisée pour bâtir des sites de commerce électronique adaptés aux besoins des utilisateurs. SETA exploite les réseaux bayésiens pour construire les modèles utilisateurs utilisés pour la recommandation et la différenciation des produits et l'adaptation du contenu. adROSA Kazienko et Adamski (2007) est un système multi-agent dédié à l'intégration des techniques du WUM et du web content mining pour réduire les données en entrée des utilisateurs afin de respecter leur vie privée.

Chaque utilisateur possède un niveau de connaissances, des capacités et des préférences particuliers. Ainsi, on a besoin de construire des systèmes adaptés à chaque type d'utilisateur ou groupe d'utilisateurs possédant des caractéristiques communes. Pour pouvoir contourner les utilisateurs hétérogènes caractérisés par leurs besoins et préférences instables, nous proposons de combiner plusieurs techniques de WUM. Ceci nous permettra de comprendre différentes caractéristiques des comportements des utilisateurs du web et par la suite découvrir davantage des connaissances significatives, pertinentes et concises, révélées sous forme de modèles. Ces modèles peuvent, par la suite être utilisés par le module de personnalisation pour adapter le comportement du système de façon appropriée.

La personnalisation web peut tirer profit du paradigme agent Mobasher (2004). Les caractéristiques de coopération, d'autonomie et l'aptitude à l'apprentissage font des agents les mieux placés pour représenter et comprendre les utilisateurs. En fait, un système multi-agent facilite l'intégration de différentes méthodes de WUM et permet la vérification et la mise à jour automatique des connaissances découvertes grâce à ces méthodes. Toutes ces raisons nous conduits à la conception de notre système multi-agent du WUM pour la personnalisation du web.

3 Conception d'un système multi-agent du WUM pour la personnalisation du web

Dans cette section, nous présentons notre méthodologie et notre politique de personnalisation adoptée. Nous décrivons également les principaux modules et agents de notre système et nous précisons leurs rôles et leurs interactions.

3.1 La méthodologie du Web Usage Mining

Notre méthodologie consiste en l'utilisation de plusieurs techniques de WUM. Pour la modélisation de l'utilisateur, nous utilisons un algorithme de clustering dynamique basé sur le modèle FM (feature Matrix) de clusters. Ce modèle permet aussi la classification en ligne des nouvelles sessions utilisateurs capturées Shahabi et Banaei-Kashsani (2003). Le modèle FM est approprié pour associer en temps réel les sessions aux clusters pré-générés. Pour construire les groupes de sessions, nous calculons des variables caractéristiques pour toutes les sessions déjà associées à des clusters en utilisant les valeurs d'une session virtuelle appelée cluster centroide ou modèle de clusters. Le résultat final de ces calculs est un ensemble de matrices de caractéristiques qui constituent le modèle FM de clusters

$$C^{fm} = \{M^{F1}, M^{F2}, \ldots, M^{Fm}\}$$

La matrice des caractéristiques du cluster est calculée pour chaque caractéristique comme suit :

$$M^F = 1/N \sum_{i=1}^{N} M_i^F$$

où N est la cardinalité du cluster C. Cette fonction est aussi appliquée, de façon incrémentale pour mettre à jour et actualiser le modèle de cluster, quand une nouvelle session Aj rejoint le cluster crée. Ce processus incrémental correspond au clustering dynamique.

$$M^F = 1/(N+1)\ (N*M^F*M_j^F)$$

Nous avons aussi adopté la nouvelle mesure de similarité définie par Shahabi et al. et basée sur le modèle FM Shahabi et Banaei-Kashsani (2003). Cette mesure est une variante de PED (Pure Euclidean Distance) appelée PPED (Projected Pure Euclidean Distance) pour éliminer le problème de la surestimation et pour réduire la complexité liée au temps des mesures de similarité. Elle est particulièrement efficace pour associer d'une manière précise des navigations partielles aux clusters, en temps réel.

$\rightarrow \quad \rightarrow$

Supposons que A et B sont deux vecteurs de caractéristiques du même type, appartenant respectivement, à une session et à un modèle de clusters. Chaque vecteur est composé de N composants. La dissemblance entre les deux vecteurs est estimé comme la somme de la PED entre les deux vecteurs et la projection du deuxième vecteur dans le plan des coordonnées où le premier vecteur n'a pas des composants nuls.

$$PPED(\rightarrow A, \rightarrow B) = \left(\sum_{i=1,\ ai\neq 0}^{N} (a_i - b_i)2\right)^{1/2}$$

Où PPED € [0, ∞] et elle est non commutative.

En comparant le premier vecteur avec le projeté du deuxième, la session et le cluster sont comparés en se basant seulement sur les segments qui existent dans la session et non pas

dans la base toute entière. Jusque là, la partie des segments non couverte dans la session est exclue de la comparaison pour éviter la surestimation.

En outre, pour chaque groupe de session, nous voulons déterminer les épisodes fréquents en utilisant les motifs séquentiels. Dans notre méthodologie, nous utilisons l'algorithme PSP Masseglia et al (1999). Il est basé sur le même principe général que l'algorithme GSP Srikant et Agrawal (1996) mais il utilise une structure de données arborescente améliorée pour hiérarchiser les séquences candidates.

Présentation et description du système modulaire proposé

Notre système de WUM pour la personnalisation du web se compose de cinq modules : le module de collecte de données, le module de prétraitement, le module de fouille de données, le module d'analyse et d'évaluation et enfin le module de personnalisation (Voir figure FIG 1)

- Le module de collecte de données : il est chargé de collecter les données d'usage à partir du web (fichiers logs, cookies, enregistrements,…) concernant les habitudes et les comportements de navigation des utilisateurs sur les sites web. Il pourra aussi collecter des données auprès de l'utilisateur. Une fois les données collectées, ce module les transmettra au le module de prétraitement.
- Le module de prétraitement : il récupère les données d'usage disponibles auprès du module de collecte de données avec lesquelles il construit les profils individuels pour chaque utilisateur (identificateur et sessions). Les données seront par la suite structurées selon un modèle bien précis et sauvegardées dans la base de données.
- Le module de fouille de données : c'est le module principal, il est déterminant pour les actions de personnalisation à entreprendre. Il se base sur les profils individuels pour pouvoir les analyser, les classer en construisant ainsi les modèles de profils génériques, en déduire des règles,… Toutes les connaissances issues de ce module seront sauvegardées dans la base de connaissances.
- Le module d'analyse et d'évaluation des connaissances : il récupère les connaissances issues de la fouille de données pour juger leur pertinence et leur aptitude à contribuer à la personnalisation et ce en mettant à profit des mesures de pertinence.
- Le module de personnalisation web : il est constitué d'un ensemble de fonctions de personnalisation. Ces fonctions ont pour objectif d'ajouter de la valeur à l'expérience de navigation des utilisateurs d'un site Web en rendant facile leurs interactions avec ce site et en économisant leur temps. Ceci dans le but de les fidéliser.

Chaque module est responsable de la réalisation d'une phase bien particulière de processus du WUM. Dans chaque phase, un ensemble d'agents collaborent pour la réalisation de l'objectif de cette phase. Le module de collecte de données fait appel à deux agents qui collectent les données en ligne de deux cotés : l'agent données utilisateur qui collecte les données du coté utilisateur et l'agent Interface qui collecte les données du coté serveur web. Une fois les données récupérées, le module de prétraitement met en coopération trois agents : l'agent filtrage, l'agent identificateur de sessions et l'agent données manquantes pour le nettoyage et la structuration des données d'usage collectées. Une fois les données structurées, le module de fouille de données sollicite trois agents qui encapsulent chacun une méthode de fouille, à savoir l'agent clustering, l'agent classification et l'agent motifs séquentiels. Les résultats sont des règles et modèles de connaissances qui seront exploitables par le module de personnalisation, constitué de deux agents : l'agent personnalisation et l'agent interface.

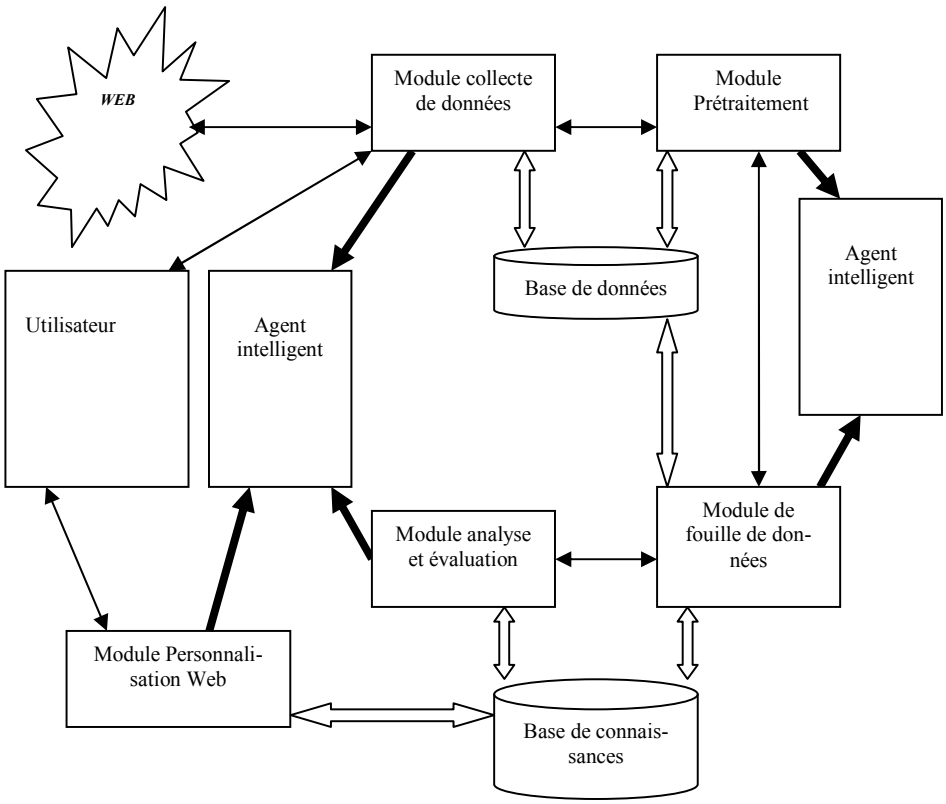


FIG. 1 – *le système modulaire du WUM pour la personnalisation du web*

3.2 Rôle des agents et interactions entre eux

Les agents proposés sont scindés en deux grands groupes : ceux qui opérèrent en ligne et ceux qui opérèrent hors ligne.

3.2.1 L'activité en ligne

Elle est assurée par les agents suivants :

- L'agent Interface : il a pour mission d'entretenir le dialogue de l'utilisateur avec le web. Il assiste également l'utilisateur dans ses interactions avec le web. Il observe les comportements de chaque utilisateur pour apprendre à partir de leurs habitudes et comportements. Une fois le processus de WUM achevé et les taches de personnalisation choisies, il affiche les résultats demandés et recommandés.
- L'agent données utilisateurs : il collecte les données utilisateurs à partir des cookies. Cet agent permet d'identifier d'une manière unique l'utilisateur en classant par sujet les données et les enregistre dans la base de données.
- L'agent classification : il classe la session utilisateur active parmi un des groupes d'utilisateurs déjà trouvés par l'agent

- l'agent évaluateur : il a deux rôles, la validation des règles découvertes et la visualisation de tous les résultats. Pour chaque règle découverte, l'agent évaluateur calcule la confiance et le support pour juger sa pertinence. La tâche de validation Kamdar et Joshi (2000) consiste en l'élimination des règles inutiles et dépourvues de sens et l'extraction des règles significatives. Ces règles seront par la suite utilisées dans les tâches de personnalisation.
- L'agent décideur : il prend les décisions quand aux fonctions de personnalisation nécessaires à entreprendre pour satisfaire les besoins de l'utilisateur. Une fois la fonction de personnalisation choisie, il active l'agent de personnalisation qui effectuera le travail demandé par cette fonction.
- L'agent personnalisation : il va entreprendre les actions de personnalisation adaptées à l'utilisateur de la session active en se basant sur le groupe de l'utilisateur trouvé. Ces actions peuvent être simples comme une assistance ordinaire par proposition d'adresses URLs susceptibles de l'intéresser, des messages de bienvenue, une mémorisation ou bien des actions plus avancées comme l'adaptation de la structure du site à l'utilisateur courant, l'exécution des procédures sans son intervention comme une négociation de prix par exemple.

3.2.2 L'activité hors ligne

Elle est assurée par les agents suivants :

- L'agent filtrage : il nettoie les données web collectées en supprimant toutes les données redondantes et inutiles. On a appliqué la méthode de Cooley et on a considéré que les entrées des fichiers logs qui représentent des images, des sons, des vidéos, autres sites web, des scripts CGI comme des données dépourvues de sens. Nous avons aussi filtré et éliminé les requêtes qui n'ont pas abouties et les requêtes des agents de recherche reconnues par le champ UserAgent du fichier log.
- L'agent identificateur de sessions : il identifie les sessions et les utilisateurs. Les utilisateurs peuvent être identifiés à travers les cookies Cooley et al (1999a) et le couple (IP, UserAgent), tandis que pour l'identification des sessions, on a utilisé une méthode basée sur le temps entre deux visites. Un ensemble de pages visitées par un utilisateur spécifique est considéré comme une seule session utilisateur si cela se produit dans un intervalle de temps inférieur ou égal à 30 minutes Cooley et al (1999a).
- L'agent données manquantes : du fait des problèmes causés par les caches et les proxys qui entrainent la perte de certaines données lors du nettoyage, cet agent aura pour rôle la détermination des données manquantes en déroulant un algorithme de détection par les fichiers log (logparser) Murgue (2006).
- L'agent clustering : il a pour rôle de grouper les utilisateurs ayant des comportements de navigation similaires. Cet agent exécute l'algorithme de clustering dynamique présenté dans la section 3.1
- L'agent motif séquentiel : il découvre les épisodes fréquents parmi les sessions utilisateurs déjà identifiées. Il facilite l'identification des pages liées et il permet la génération des patrons sous forme de règles de navigation.

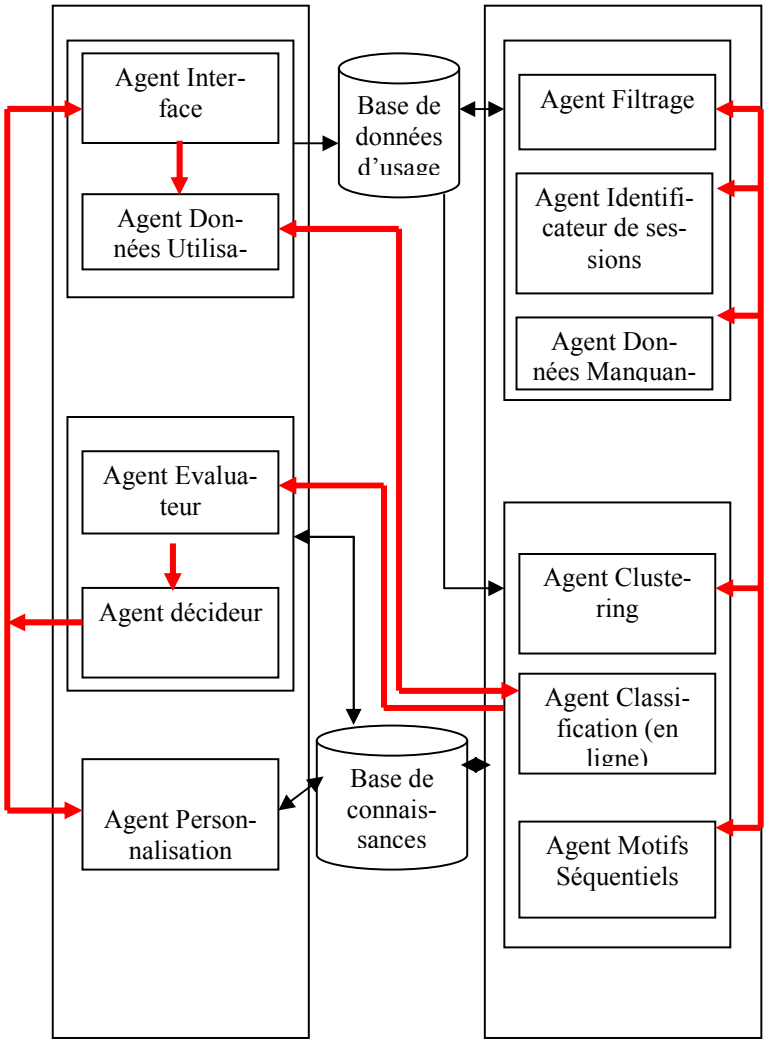


FIG. 2 – *le système multi-agent du WUM pour la personnalisation du web*

3.2.3 Les interactions entre les agents

Le but de notre système est la découverte des connaissances opérationnelles pour automatiser le processus de personnalisation. Pour réaliser un tel objectif, les agents coopèrent et se coordonnent par le biais de messages. Ces messages sont des messages informatifs envoyés soit pour activer d'autres agents, soit pour donner des informations. L'agent interface et l'agent données utilisateurs sont deux agents mobiles qui s'exécutent en parallèle. Ils sont créés à chaque fois qu'il y a un nouvel utilisateur. L'agent interface capture les comportements de navigation de l'utilisateur courant et informe l'agent données utilisateurs et l'agent classification en ligne de ces comportements (A1 et A2 in FIG3). L'agent données utilisateurs identifie l'utilisateur courant, crée et sauvegarde sa nouvelle session de navigation. Par ailleurs, il demande (A3) à l'agent de classification en ligne de classer la session active parmi

les groupes découverts par l'agent clustering, en utilisant les nouvelles informations concernant l'utilisateur se trouvant dans sa session.

Les interactions entre les agents sont représentées dans la figure 3 en utilisant AUML (Agent-base Unified Modeling Language) qui est une variante d'UML pour modéliser les interactions entre les agents (http://www.auml.org/). AUML est un langage de modélisation des systèmes multi-agent Bauer (1999). D'une part, les agents sont actifs, ils sont capables de prendre des initiatives et peuvent contrôler la communication entre eux Bauer et al (2000). D'une autre part, les agents coopèrent et coordonnent leurs travaux pour atteindre un but commun.

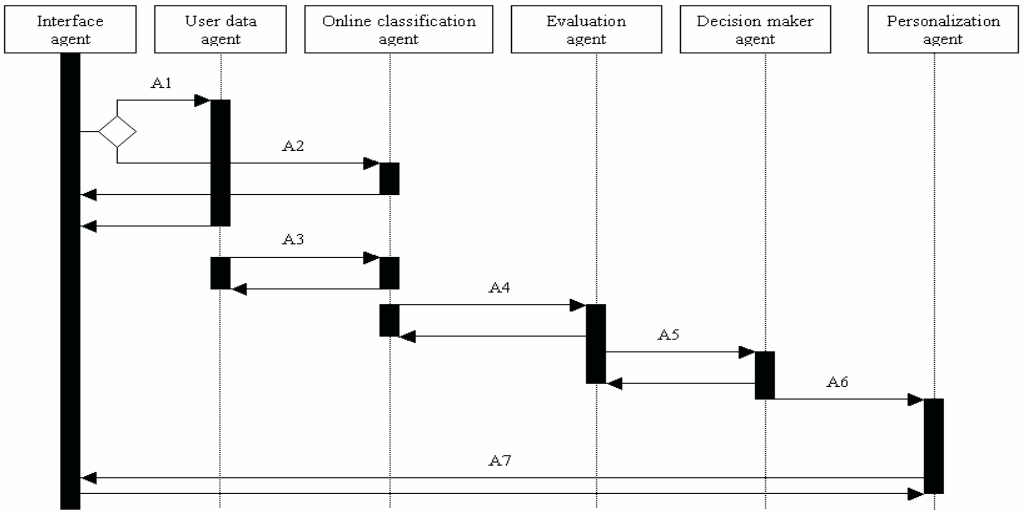


FIG. 3 – Diagramme de séquences AUML représentant les interactions entre les agents au cours du travail en ligne.

3.3 Politique adoptée de personnalisation

Le seul responsable des tâches de personnalisation est l'agent de personnalisation. Il exploite les connaissances issues des modèles des utilisateurs et des motifs séquentiels découverts précédemment. Il applique un ensemble de règles de personnalisation pour appliquer une fonction ou tache de personnalisation parmi les taches et les fonctions suivantes :

- la mémorisation des informations personnelles des utilisateurs,
- l'affichage de messages de bienvenue aux utilisateurs fidèles,
- la recommandation de liens précédemment choisis et visités par les utilisateurs du groupe auquel appartient l'utilisateur courant,
- la différenciation des objets en présentant différentes caractéristiques pour chaque objet demandé.

Cependant, la manière avec laquelle ces fonctions seront combinées pour procurer une solution de personnalisation complète, dépend de la politique de personnalisation que le propriétaire du site souhaite avoir. Du fait que nos fonctions de personnalisation sont exécutées seulement au début des sessions utilisateurs, notre personnalisation est statique.

4 Réalisation

La réalisation du PWUM décrit ci-dessus s'appuie principalement sur les deux éléments suivants :

- L'utilisation de plusieurs techniques d Web Usage Mining tel que l'algorithme de clustering dynamique basé sur le modèle FM, qui aussi permet la classification en ligne des nouvelles sessions utilisateurs capturées et l'algorithme PSP de découverte de motifs séquentiels, qui permet de déterminer les épisodes fréquents pour chaque groupe de sessions. Nous disposons pour tester nos algorithmes de données extraites du domaine du tourisme dans le cadre du projet RNTL Eiffel [quelles références ?]. Ces données représentent des contextes de navigation variés et sont utiles pour simuler divers scénarios de personnalisation.
- L'utilisation d'une plateforme multi-agents pour l'implémentation du processus d'interaction entre les différents agents du système. Les agents logiciels de PWUM ont été implémentés en utilisant une plateforme multi-agent qui s'appelle JADE (Java Agent DEvelopment Framework), Bellifemine et al (2004), Caire (2003). JADE est une plateforme publique entièrement implémentée en JAVA. Elle simplifie l'implémentation des systèmes multi-agent à travers un middleware basé sur les standards de la FIPA (Foundation for Intelligent Physical Agents). JADE permet aussi l'interopérabilité des systèmes, ainsi, les agents peuvent être distribués entre différentes machines indépendamment du système d'exploitation qu'ils utilisent. Enfin, à travers le mécanisme des messages JADE, FIPA-ACL (http://www.fipa.org/specs/fipa00026/sc00026H.html/), les agents ont l'autonomie dans la gestion des messages reçus et envoyés.

5 Conclusion

Dans cet article, nous avons présenté notre système multi-agent du WUM pour la personnalisation web et nous avons décrit les mécanismes nécessaires pour dérouler le processus WUM (Web Usage Mining) et effectuer les tâches de personnalisation.

Nos agents sont en cours d'implémentation, nous espérons que les résultats que nous allons obtenir en utilisant à la fois les systèmes multi-agents et les techniques WUM seront encourageants. Nous allons tester notre approche dans le domaine de tourisme.

Dans notre travail, nous avons utilisé des profils statiques pour modéliser les utilisateurs du web, nous espérons dans des futures directions pouvoir transformer nos profils en profils dynamiques vue que les utilisateurs sur le web changent toujours de comportements et d'habitudes. Aussi, le web est un système dynamique et complexe, il est toujours en changement perpétuel. Le fait d'utiliser des profils dynamiques nous permettra de tenir compte de tous ces changements et de mettre à jour les modèles des utilisateurs pour pouvoir les satisfaire de la manière la plus optimale que possible.

Références

Ardissono L. and Torasso P. (2000), *Dynamic user modelling in a web store shell*. In Proceedings of the 14th Conference ECAI, Berlin, Germany, pp 621-625.

Bauer B. (1999), *Extending UML for the specification of interaction protocols*. Submission for the 6th call for proposal of FIPA and revised version part of FIPA 99.

Bauer B. et al.(2000).Agent UML: a formalism for specifying multiagent interactions. AOSE:91-104.

Bellifemine F. et al. (2004), *JADE Basic Documentation: Programmer's Guide*.

Caire G. (2003), *Developing multi-agent applications with JADE: Tutorial for beginners*.

Cooley R. (2000), *Web Usage Mining: Discovery and Application of Interesting patterns from Web data.* PhD Thesis, Dept.of Computer Sciences, University of Minesota.

Cooley R. et al. (1999a), Data Preparation for mining World Wide Web browsing patterns. *Journal of Knowledge and Information System*, pp 55-32.

Girardi R. et al. (2005),WUMA-Miner: An Agent-based Design Pattern for Mining Users with similar Navigational behaviour.

Jiang Q. (2003), *Web Usage Mining: Processes and Application*, CSE 8331.

Kamdar T. and Joshi A. (2000), *On Creating Adaptive Web Sites using WebLog Mining*", Technical Report TR-CS-00-05, Department of Computer Sciences and Electrical Engineering, University of Maryland, Baltimore Country.

Kazienko P. and Adamski M. (2007). *AdROSA-Adaptive Personalization of web advertising.* Information Sciences: an International Journal, Volume 177 , Issue 11, Pages 2269-2295

MacQueen J.B. (1967), Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley, University of California Press, 1:281-297

Madria S.K. (2007), *Web Mining: A bird's Eye View*. The 8th International Conference on Web Information Systems Engineering, WISE 2007.

Masseglia F. et al. (1999), An efficient algorithm for web usage mining. *Networking and Information System Journal*.

Mobasher B. (2004), *Web Usage Mining and Personalization.* Practical Hand Book of Internet Computing, Chapman Hall & CRC Press,Baton Rouge.

Murgue T. (2005), *De l'importance du prétraitement des données pour l'utilisation de l'inférence grammaticale en Web Usage Mining*, Atelier sur la modélisation utilisateurs et personnalisation de l'interaction homme-machine (EGC'05), Paris, 2005.

Pierrakos D. et al. (2003), *Web Usage Mining as a tool for Personalization: a survey*.User Modeling and User Adapted Interaction 13, 311-372, Kluwer Academic Publishers: Netherlands.

Shahabi C. and Banaei-Kashsani F. (2003), Efficient and Anonymous Web Usage Mining for Web Personalization. *INFORMS Journal on Computing-Special Issue on data mining*, Vol 15, No.2.

Srikant R. and Agrawal R. (1996), Mining sequential patterns: Generalization and performance improvements, *In Proceeding of the 5th International Conference on Extending Database technology*, Avignon, France, page 3-17.

Symeonidis A.L. and Mitkas P.A. (2006), agent Intelligence through data mining. *the 17th European conference on Machine Learning and the 10th European Conference on Principles and Practice of Knowledge Discovery in Data base*. Berlin, Germany.

Tanasa D. (2005), *Web Usage Mining: contribution to intersites logs preprocessing and sequential pattern extraction with low support*. PhD thesis, INRIA Sophia Antipolis, AXIS project.

http://www.auml.org/

FIPA Request Interaction Protocol Specification, 2004, Available in: http://www.fipa.org/specs/fipa00026/sc00026H.html/, accessed in 22 January 2005.

Summary

This work described PWUM, a web usage mining multi agent system for web personalization. We presented our system and described the mechanism necessary for WUM (Web Usage Mining) and personalization tasks. Main algorithms of personalization has been introduced along with the details of interactions between the entities of the multi-agents system. The software agents of PWUM has been implemented using a multi-agents platform called JADE.

Description de cartes géographiques

Catherine Dominguès

Institut géographique national
Laboratoire COGIT
2/4 avenue Pasteur
94165 Saint-Mandé cedex
catherine.domingues@ign.fr
http://recherche.ign.fr/cogit

Résumé. Dans un contexte de conception de carte sur mesure, notre proposition consiste à exploiter la description qu'un utilisateur fait de sa carte pour l'aider à concevoir une carte efficace et adaptée à son besoin. Cet article présente une expérimentation mise en place auprès d'utilisateurs afin de recueillir des commentaires sur des cartes géographiques. L'exploitation de ces commentaires, guidée par des compétences en cartographie et des connaissances linguistiques, fondée sur des réécritures et des interprétations, conduit à une ébauche de description formelle de la carte.

1 Introduction

Le laboratoire COGIT de l'Institut géographique national travaille à un projet de conception de carte sur mesure. La conception d'une carte géographique est un processus qui comporte plusieurs étapes, de la spécification des objectifs de la carte jusqu'à son affichage. Il s'agit, à chaque étape, de proposer à un utilisateur néophyte une aide logicielle experte afin qu'il puisse créer une carte adaptée à son besoin, c'est-à-dire pertinente et efficace par rapport à ses objectifs et adaptée à ses goûts.

Dans ce contexte, notre proposition vise à constituer des bases de connaissances qui permettront d'exploiter la description que l'utilisateur fait de son besoin pour l'aider à concevoir sa carte dans les règles de l'art. Pour ce faire, il est nécessaire d'établir des correspondances entre la description que des utilisateurs peuvent faire de leur carte et ses paramètres de construction tels qu'ils sont définis par des experts.

Différentes actions ont été mises en place pour faire décrire des éléments caractéristiques d'une carte (Buard et Ruas, 2007) ou une carte complète (Dominguès et Bucher, 2006) par des publics variés. L'objet de cet article est de présenter un de ces travaux. Nous décrirons l'expérimentation mise en place afin de constituer un corpus de commentaires, puis l'exploitation de ce corpus qui, combinée à une méthode de construction d'une carte topographique que nous présenterons succinctement, a conduit à une succession de réécritures et interprétations des commentaires initiaux. La description formalisée d'une carte que nous ébauchons en fin d'article se fonde sur ces réécritures et interprétations.

2 Constitution du corpus

Pour constituer ce corpus de commentaires, une expérimentation a été menée auprès des chercheurs (essentiellement des géomaticiens) et stagiaires du laboratoire afin de recueillir des commentaires variés, libres et spontanés sur des échantillons cartographiques qui leur ont été proposés. Dans ce paragraphe, nous présentons cette expérience, puis les traitements appliqués aux commentaires sont résumés sur un schéma.

2.1 Mise en place de l'expérimentation

Nous avons assemblé une collection de cartes géographiques : des cartes topographiques et quelques cartes thématiques[1], cartes réelles ou échantillons construits ad hoc, françaises ou étrangères, à différentes échelles, couvrant des types de zones différents. Des échantillons de ces cartes, tous de même taille, ont été scannés et assemblés en un diaporama de 40 diapositives[*]. Les 21 sujets ont été rassemblés dans une salle où le diaporama leur a été présenté à cadence régulière (une carte à commenter toutes les 20 secondes[*]). Chaque sujet disposait d'un formulaire papier sur lequel il avait pour consigne de noter, sous forme textuelle, tous les commentaires que lui inspirait l'échantillon présenté. Aucune restriction thématique n'était apportée aux commentaires qui pouvaient concerner les informations délivrées par la carte, ses caractéristiques techniques, ses qualités esthétiques, des associations d'idées, etc. De même, le format syntaxique des commentaires était complètement libre : mots isolés, groupes nominaux (GN), phrases sur le modèle <*sujet*> <*verbe*> <*complément*>, etc. La figure 1 montre un extrait du formulaire de saisie et des exemples de commentaires.

D4 : *pâle, illisible, triste* .....
pas attrayante .....
contraste faible .....
C9 : *jolie mais détails illisibles*.....
pas assez contrastée.....
E44 : *classique pour routier*.....
couleurs bien contrastées.....
E37 : W*aouh ! Trop beau* .....
géol ? non, mais on dirait .....

FIG. 1 – *Extrait du formulaire de saisie des commentaires*

1 Une carte topographique représente essentiellement des éléments topographiques, alors qu'une carte thématique représente, sur un fond repère le plus souvent topographique, des phénomènes localisables de toute nature, qualitatifs ou quantitatifs. (CFC, 1990).

[*] Le test a été calibré au préalable auprès de 3 sujets volontaires : une temporisation de 20 secondes laissait aux sujets assez de temps pour écrire tout en restant spontanés ; 40 diapositives pour une quinzaine de minutes de test (3 diapositives ont d'abord été présentées pour que les sujets se familiarisent au rythme) ont paru constituer une durée maximale qui évite la lassitude des sujets.

2.2 Pré-traitements du corpus

Les commentaires se présentent sous forme de lignes écrites sur papier. Ce sont le plus souvent des groupes nominaux (GN), de patron syntaxique <*nom*> <*modifieur*> où le modifieur peut être plus ou moins complexe. Ces commentaires doivent être saisis sur support informatique et mis en forme pour constituer un corpus homogène et exploitable avec des outils informatiques. Les traitements nécessaires sont fondés sur des critères lexicaux et syntaxiques. Ils ne sont pas décrits dans cet article mais résumés et illustrés à chaque étape par des exemples du corpus, dans la figure 2.

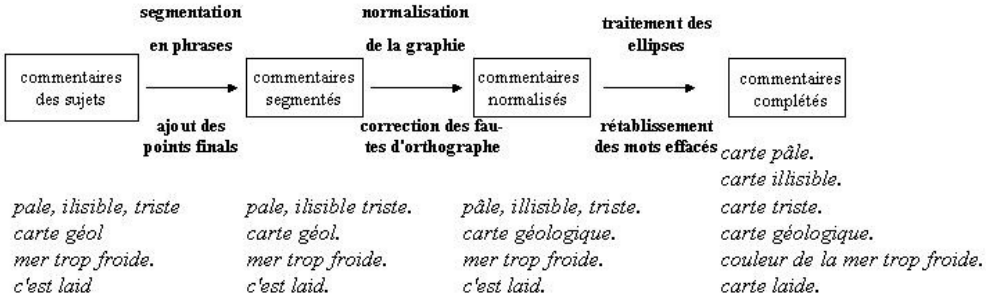


Fig. 2 – Traitements et réécritures appliqués aux commentaires

Les commentaires initiaux formaient 1 290 lignes ; après la mise en forme, en particulier le traitement des ellipses, le corpus en contient environ 1 400. Le tableau 1 présente une des cartes et certains des commentaires (après mise en forme) qui s'y rapportent.

carte A4

Sujets	Commentaires après réécritures
6	*carte brillante*
12	*carte chaude*
15	*carte épicée*
8	*beau contraste*
5	*contraste fort*
12	*trop de contrastes*
18	*carte trop détaillée*
1	*carte moche*
9	*carte lumineuse*
6	*carte riche*
5	*carte vivante*
7	*couleurs mauvaises*
2	*feu*
19	*Maghreb*
9	*soleil*
9	*vacances*

21	*que sont les choses jaunes ?*

TAB. 1 – *Carte A4 et commentaires s'y rapportant*

L'examen du corpus de commentaires, dont le tableau 1 donne un extrait, appelle différentes remarques sur l'identification du thème sur lequel porte le commentaire. Ces remarques vont être discutées dans le paragraphe suivant.

2.3 Identification du thème sur lequel porte le commentaire

Pour être exploitables, les commentaires doivent être rattachés à un concept cartographique ou à une propriété de la carte ; tous ne sont donc pas exploitables tels qu'ils sont formulés.

Certains sont explicites ; ils concernent un thème (l'aspect général de la carte, ses couleurs, les contrastes) facilement identifiable par le vocabulaire de l'énoncé lui-même. Les exemples de ce type sont : *carte fade* ou *couleurs trop claires* ou *trop de détails*.

Les énoncés qui décrivent des associations d'idées : *Far-West*, *lacs*, *grand espace*, *Las Vegas.*, *carte d'un pays pauvre* relèvent aussi de ce cas. Ils ont le plus souvent la forme d'un nom seul, nom commun ou nom propre, simple ou composé. Ils ne sont pas directement rattachables à un objet géographique représenté sur la carte ou à un concept cartographique mais sont intéressants par les associations d'idées qu'ils mettent en place.

D'autres énoncés, malgré les mises en forme et réécritures évoquées au paragraphe 2.2, ne présentent pas une formulation satisfaisante parce qu'elle ne rend pas possibles leur comparaison et leur regroupement sémantiques, guidés par le vocabulaire. Une association fondée sur le vocabulaire est impossible parce que l'idée contenue dans le commentaire nécessite d'être interprétée pour être rattachée à un concept pertinent de la cartographie. Une reformulation qui utilise des compétences en cartographie s'avère nécessaire. Ces interprétations et reformulations sont manuelles et dépendent des concepts retenus pour décrire les règles de construction de la carte. Par exemple, à la fin du tableau 1, *que sont les choses jaunes ?* est un commentaire qui nécessite une interprétation supplémentaire pour être exploitable. Une interprétation plausible serait : puisque les symboles jaunes ne sont pas facilement interprétables, c'est que la couleur jaune n'est pas adaptée à l'objet géographique qu'elle représente (en l'occurrence un terrain de sport).

Les sujets ayant toute liberté dans la rédaction de leurs commentaires, ceux-ci concernent de multiples aspects de la conception d'une carte : les propriétés visuelles de la carte, ses caractéristiques techniques, les associations d'idées qu'elle a engendrées, etc. Trois types de commentaires ont été identifiés qui conduisent à répartir les commentaires en trois classes :

- ceux concernant l'apparence générale de la carte, ses caractéristiques globales, visuelles et l'expression du ressenti de l'utilisateur par rapport à ces caractéristiques ;
- ceux concernant l'application des règles de construction des cartes, en rapport avec la sémiologie graphique[2] et l'utilisation des couleurs ;
- ceux concernant les associations d'idées induites par la carte.

[2] La sémiologie graphique est l'ensemble des règles qui régissent la construction d'un système de signes permettant la traduction graphique d'une information. Elle a été notamment étudiée par Jacques Bertin (Bertin, 1998).

Dans la suite de l'exposé, les commentaires mis en forme, éventuellement interprétés et reformulés, seront désignés par le mot énoncé.

2.4 Classement et organisation des énoncés

Les énoncés ont été classés en utilisant les critères définis précédemment. Puis dans chaque classe, les énoncés ont été étudiés selon des critères lexicaux afin de préciser le thème auquel ils se rapportent.

Certains énoncés ont été identifiés comme appartenant à deux classes parce qu'ils sont ambigus ou parce qu'ils relèvent d'idées pouvant appartenir à deux types d'énoncés, par exemple : *carte riche*. Celui-ci peut évoquer l'aspect général de la carte, et en particulier la richesse, la variété de ses coloris, et peut aussi faire référence à la quantité d'informations représentée, ce qui relève des règles de sémiologie graphique adoptées pour la construction de la carte.

Le tableau 2 montre le résultat de ce classement pour la carte dessinée dans le tableau 1. Les énoncés de la deuxième colonne commentent la carte en général à travers le ressenti de l'utilisateur : *carte épicée*. Les énoncés de la troisième colonne font référence à l'application de règles de sémiologie graphique pertinentes en cartographie : les contrastes (*contraste fort*) ou le choix des couleurs (*couleurs mauvaises*). Les énoncés de la dernière colonne (*feu*, *sécheresse*) relèvent des associations d'idées (les GN ont été lemmatisés). Un énoncé peut figurer dans plusieurs colonnes.

Commentaires	**Aspect général, ressenti du sujet**	**Règles de construction, sémiologie graphique, couleurs**	**Associations d'idées**
carte brillante	*carte brillante*		
carte chaude	*carte chaude*		
carte épicée	*carte épicée*		*carte épicée*
beau contraste		*beau contraste*	
contraste fort		*contraste fort*	
trop de contrastes		*trop de contrastes*	
carte trop détaillée		*carte trop détaillée*	
carte moche	*carte moche*		
carte lumineuse	*carte lumineuse*		
carte riche	*carte riche*	*carte riche*	
carte vivante	*carte vivante*		
couleurs mauvaises		*couleurs mauvaises*	
que sont les choses jaunes ?		*couleur jaune mal adaptée aux équipements sportifs*	
feu			*feu*
Maghreb			*Maghreb*
soleil			*soleil*
vacances			*vacances*

TAB. 2 – *Répartition des énoncés en trois classes.*

Le regroupement des énoncés en trois classes constitue une première étape vers la description formelle d'une carte. Avant de conceptualiser le contenu de ces classes en rattachant l'énoncé à un concept cartographique ou à une propriété de la carte, nous présentons succinctement, dans le paragraphe suivant, une méthode de construction d'une carte topographique sur laquelle nous pourrons articuler la description formelle d'une carte.

3 Démarche de conception et réalisation d'une carte

La cartographie a pour objectif de donner une représentation graphique, sous forme de carte, d'informations géographiques. Cette représentation relève à la fois d'une pratique artistique et d'une démarche scientifique spécifique (Bertin, 1998), (Cuenin, 1972). Comme l'indiquent Zanin et Trémélo (2003), "Cette démarche, pour être efficace et la plus objective possible, exige l'application d'un certain nombre de principes et l'apprentissage du langage graphique que l'on désigne sous le terme générique de "sémiologie graphique".

Ces auteurs décrivent la démarche de conception d'une carte thématique. Cette démarche est simplifiée pour la réalisation d'une carte topographique et repose sur les étapes suivantes :

1. "identifier l'objectif de la carte". Il s'agit d'initialiser les paramètres suivants :
   - les informations, le message à transmettre,
   - le type de carte à construire : carte illustration d'un texte, carte de manuel, carte d'atlas, etc,
   - le thème majeur de la problématique de la carte, identifié d'après le message à transmettre et les données géographiques disponibles,
2. "identifier la cible : le public, le support, les conditions d'utilisation". Ces paramètres peuvent prendre les valeurs suivantes :
   - le public : lecteurs de presse, professionnels (agriculteurs, pompiers), étudiants, collégiens, etc,
   - le support : rapport d'étude, site Internet, presse (généraliste, spécialisée, grand public, …), livre pour enfant, atlas, manuel didactique, etc,
   - les conditions d'utilisation (support de travail, utilisation familiale, en plein air) et les conditions d'éclairage (éclairage naturel, sources lumineuses artificielles),
   - les dimensions de la carte,
   - le type de représentation : couleur ou noir et blanc,
3. "identifier l'information à cartographier" : quelles données retenir pour répondre à la problématique et comment les traiter pour mettre en valeur et traduire correctement leurs caractéristiques essentielles ;
4. "identifier la figuration de l'implantation" i.e. le support graphique de l'information. Il faut mettre en place le support graphique le mieux adapté au public et à l'information à transmettre. Pour cela il faut définir :
   - le "niveau géographique" de la carte : monde, continent, état, ville, …
   - le type de projection le mieux adapté,
   - l'échelle,
   - le niveau de détail souhaité qui se traduit par le choix de la généralisation[3],

[3] La généralisation est l'adaptation des données qualitatives et quantitatives, par allègement du nombre de détails et simplification caractérisée des formes des tracés, en vue de l'établissement d'une carte répondant à des conditions déterminées. (Comité français de cartographie, 1990).

5. "identifier et organiser la figuration graphique de l'information". Il faut choisir les variables visuelles[4] pertinentes en fonction des données à représenter et des règles de sémiologie graphique (Bertin, 1998), (Cuenin, 1972), (Zanin et Trémélo, 2003).

Cette démarche de conception d'une carte s'appuie sur des concepts et des objets définis par des experts. La description d'une carte telle qu'elle apparaît au travers des énoncés des sujets n'est ni conceptualisée ni aussi complète. Elle ne renseigne pas la totalité des paramètres définis dans la méthode exposée ici. Dans le paragraphe suivant, nous nous proposons de formaliser les énoncés en utilisant les concepts définis par les experts.

4 Formalisation des énoncés

Les énoncés ont été précédemment regroupés en trois classes (paragraphe 2.4). Nous étudions une formalisation des énoncés dans chaque classe. Une carte étant commentée par des énoncés appartenant aux trois classes, elle sera formellement décrite par l'ensemble des transcriptions des énoncés.

4.1 Enoncés concernant l'apparence générale de la carte

Ces énoncés décrivant l'aspect général de la carte à travers le ressenti exprimé par les sujets sont, en général, de la forme *: carte <modifieur>*.

Les modifieurs sont des adjectifs, éventuellement précédés d'un adverbe permettant de marquer des différences d'intensité ; 79 adjectifs ont été recensés :

agréable + agressive + attrayante + belle + bizarre + brillante + calme + chouette + claire + clinquante + complexe + contrastée + démodée + désagréable + détaillée + douce + efficace + élégante + enfantine + ennuyeuse + ensoleillée + épicée + équilibrée + étudiée + exotique + expressive + fade + fatiguée + figurative + floue + forte + froide + gaie + glaciale + harmonieuse + homogène + illisible + joyeuse + laide + lassante + légère + lisible + lourde + lumineuse + luxuriante + marrante + mignonne + moche + nette + neutre + normale + nulle + originale + osée + pâle + pâlotte + pétante + précise + propre + psychédélique + rangée + ratée + riche + saturée + sérieuse + simple + sobre + sombre + structurée + sympa + traditionnelle + triste + vieillie + vide + vilaine + vivante + vive + vulgaire + zen

Parmi ces 79 adjectifs, 10 ont été conservés comme descripteurs de base de l'ensemble des modifieurs, les autres adjectifs s'exprimant en fonction de cette base minimale. Cette transcription s'effectue manuellement sur des critères de synonymie ou de proximité sémantique dans le contexte de la cartographie. La grammaire de description d'un commentaire de cette classe peut donc s'écrire de la manière suivante (*E* désigne le mot vide) :

énoncé =:	*carte Modifieur*
Modifieur =:	*A (AdvInt+E) | Modifieur ; Modifieur*
AdvInt=:	*peu+pas+très+trop*
A =:	*belle + chaude + contrastée + équilibrée + lumineuse + originale + pastel + précise + réaliste + sobre*

Avec cette grammaire, la formulation de :

- *carte agressive* devient *carte contrastée trop ; équilibrée pas ; sobre pas*
- *carte agréable* devient *carte belle ; équilibrée ;*

[4] Les variables visuelles sont au nombre de sept dont la couleur, la valeur, la forme et la taille.

- *carte riche* devient *carte sobre pas* ; *complexe*
  (par rapport à l'aspect général de la carte)

4.2 Enoncés concernant l'application des règles de construction

Les paramètres, à la fois utilisés dans la méthode des experts et identifiés dans les énoncés produits par les sujets, sont les suivants :

- le type de carte ;
- le public auquel est destinée la carte ;
- le thème dominant, et plus particulièrement la désignation du thème dominant, la couleur associée, son caractère discriminant, le contraste qu'elle forme avec les autres symboles de la carte ;
- la couleur, d'abord l'appréciation de la composition des couleurs (intuitivement ou par rapport aux règles de sémiologie graphique) et aussi les concepts de dominante et de contraste ;
- la densité ;
- l'échelle.

Dans les énoncés, ces paramètres sont évalués sous la forme de GN : *couleurs contrastées*, *gros contraste*, *carte trop détaillée*. Pour formaliser l'idée contenue dans le commentaire, il faut indiquer le nom du paramètre pertinent et son modifieur. Selon le cas, il faut un ou plusieurs niveaux de noms pour préciser le concept comme pour : *couleur dominante jaune <modifieur>*. Les modifieurs nécessaires pour exprimer tous les aspects pertinents du concept sont aussi plus variés que les adjectifs associés à *carte* dans le paragraphe précédent.

Pour rendre compte de ces différents paramètres et formaliser complètement les énoncés, il faudrait définir pour chaque paramètre retenu l'ensemble des valeurs qu'il peut prendre. Ce travail n'est pas encore terminé. En particulier, les valeurs que peuvent prendre le type de carte et le public concerné doivent être complétées et organisées pour que les différences de vocabulaire puissent renvoyer à des différences de cartes substantielles. En revanche, d'autres paramètres sont complètement définis comme les couleurs et leurs différentes intensités (ou valeurs).

Ces couleurs (notées *NomCouleur*) sont celles du cercle chromatique défini dans (Chesneau, 2005) pour étudier les contrastes colorés : 14 couleurs (nommées couleurCercle dans la grammaire) de 7 intensités chacune (du 1 : très clair au 7 : très foncé) auxquelles s'ajoutent un blanc, un noir et une gamme de 7 gris, plus des couleurs grisées, nommées couleurGrisée, (les 14 couleurs précédentes désaturées avec du gris en 4 paliers), soit au total 163 tonalités.

La grammaire de codage de ces informations pourrait être la suivante (les non terminaux *énoncéTypeDeCarte*, *énoncéPublicConcerné* et *énoncéEchelle* restent à préciser) :

AdjEvaluationDensité =	*faible* | *adaptée* | *forte*
AdjEvaluation =:	*mauvais* | *acceptable* | *réussi*
NomCouleur =:	*noir*| *blanc*| *couleurCercle intensitéCouleur*| *grisColoré intensitéGris*| *couleurGrisée intensitéGris*
couleurCercle =:	*pourpre*|*rougeViolet*|*rouge*|*orange*|*jauneOrange*|*jaune*| *vertJaune*|*vert*|*vertBleu*|*bleu*|*bleuViolet*|*violet*|*ocre*|*marron*|*gris*
couleurGrisée =:	*pourpreGrisé*| *rougeVioletGrisé*| *rougeGrisé*| *orangeGrisé*| *jauneOrangeGrisé*| *jauneGrisé*| *vertJauneGrisé*| *vertGrisé*| *vertBleuGrisé*| *bleuGrisé*| *bleuVioletGrisé*| *violetGrisé*| *ocreGrisé*| *marronGrisé*

intensitéCouleur =: *1| 2| 3| 4| 5| 6| 7*
intensitéGris =: *1| 2| 3| 4*
énoncé = *énoncéTypeDeCarte | énoncéPublicConcerné | énoncéThème | énoncéCouleur | énoncéDensité | énoncéEchelle*
énoncéThème =: *thème (couleur NomCouleur AdjEvaluation + E) (contraste AdjEvaluation + E) (discrimination AdjEvaluation + E)*
énoncéCouleur =: *couleur (dominante NomCouleur AdjEvaluation+ E) (contraste AdjEvaluation + E) (NomCouleur AdjEvaluation + E)*
énoncéDensité =: *densité AdjEvaluationDensité*

Avec cette grammaire, la formulation de :

- *couleurs mauvaises* devient *couleur composition à éviter*
- *beau contraste* *devient* *couleur contraste réussi*
- *carte riche* devient *densité forte*
  (par rapport à la quantité d'informations représentée)

4.3 Enoncés concernant les associations d'idées induites par la carte

Les associations d'idées proposent différents types de références :

- référence à un lieu :
  - par un toponyme qui n'appartient pas à la carte commentée : *Maghreb*, *Las Vegas,*
  - par un toponyme imaginaire : *Atlantide*,
  - par une expression désignant des objets géographiques qui ne figurent pas sur la carte commentée : *oasis*, *désert, forêt, fjord, goulag, ville engloutie*,
- référence à la nourriture : *aubergine, cerise, chocolat, glace, gouda, tomate, vin*,
- référence à un nom non concret : *tristesse, tranquillité, art moderne, pollution, réchauffement planétaire, ambiance glacée, fin du monde*
- référence à un nom concret : *vache, feu de forêt*.

Une exploitation immédiate de ces associations d'idées consisterait à faire correspondre le type de carte avec un énoncé pertinent, par exemple :

- une carte touristique sur la route des vins avec une carte associée à *vin* ;
- une carte touristique pour une zone littorale ou une station balnéaire avec une carte associée à *vacances* ou *soleil* ;
- une carte dont le principal thème à traiter est l'environnement avec une carte associée, selon l'objectif de la carte, à *pollution*, *réchauffement planétaire* ou *ambiance glacée*.

5 Conclusions et perspectives

L'objectif de ce travail était d'ébaucher une description formalisée d'une carte topographique, qui s'appuie à la fois sur la démarche des experts et le ressenti des sujets- concepteurs-utilisateurs potentiels de cartes.

L'étude de la démarche des experts a permis de définir les paramètres nécessaires à l'élaboration de la carte, en référence aux règles de sémiologie graphique utilisées en cartographie.

Nous avons montré que les commentaires des utilisateurs couvrent un champ sémantique moins précis et plus large que celui des experts, et intègrent en particulier le ressenti des utilisateurs. Ce ressenti qui se traduit le plus souvent par un commentaire sur l'aspect général de la carte n'est pas étudié dans la littérature didactique mais est important pour l'utilisateur qui l'exprime spontanément (ce type de commentaires constitue en effet le plus gros effectif dans l'expérimentation). La définition, ébauchée ici, d'un ensemble de descripteurs de base devrait permettre de décrire le besoin et le désir de l'utilisateur pour toutes les cartes qu'il souhaiterait construire.

Cependant, la description de l'ensemble des paramètres identifiés dans les commentaires doit être affinée. En effet, certains paramètres ne sont pas évoqués spontanément par les utilisateurs et nécessitent pourtant d'être renseignés explicitement avant de pouvoir réaliser la carte. Des règles d'inférence (à définir) pourront déduire de certains commentaires des valeurs de paramètres nécessaires à la construction de la carte.

La sélection d'une carte sur internet par un autre critère que son titre est difficile, voire impossible. Notre laboratoire dispose aussi d'un corpus de cartes. Une fois la description formelle complète, nous souhaitons l'utiliser pour décrire formellement les cartes de ce corpus et en faciliter la recherche.

Références

Bertin, J. (1998), *Sémiologie graphique: les diagrammes, les réseaux, les cartes.* Première édition en 1967 puis en 1973, 1998. Paris. Editions de l'EHESS.

Buard, E., A. Ruas (2007). Evaluation of colour contrasts by means of expert knowledge for on-demand mapping, *23th ICA Conference*. Moscow.

Chesneau, E. (2005). Propositions méthodologiques pour l'amélioration automatique des contrastes de couleur dans les cartes de risque, *Actes du colloque SAGEO (SAGEO'2005)*. Avignon.

Cuenin, R. (1972). Cartographie générale (tome 1) Notions générales et principes d'élaboration. p. 109-179. Paris. Editions Eyrolles.

Comité Français de Cartographie (1990). *Glossaire de cartographie*. Bulletin n°123-124, mars-juin 1990. Paris

Dominguès, C., B. Bucher (2006), Application d'aide à la conception de légendes, *Actes du colloque SAGEO (SAGEO'2005)*. Strasbourg.

Zanin, C., ML. Trémélo (2003). *Savoir faire une carte*. Paris. Belin.

Summary

In a context of mapping on demand, our proposal consists of exploiting the user map description to help him to realize an efficient and relevant map. This paper introduces a users survey to gather comments about maps. The commentaries exploiting is guided by cartographical competences and linguistic knowledge, dealing with rewriting and interpreting. It leads to a draft map with formalized description.